Nonlinear Dynamical Analysis and Predictive Coding of Speech

A Thesis Submitted
in Partial Fulfilment of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

by

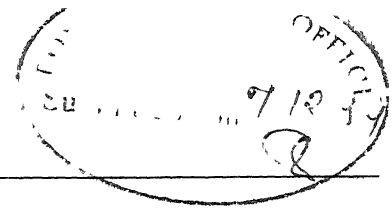
ARUN KUMAR

to the

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY KANPUR

December, 1994.

Certificate



It is certified that the work contained in the thesis entitled "Nonlinear Dynamical Analysis and Predictive Coding of Speech", by Arun Kumar, has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

S K Mullick
Professor
E E Department
I I T, Kanpur

December, 1994

To
my parents

Acknowledgements

I wish to express my deep sense of gratitude to my supervisor Professor S K. Mullick for his guidance, advice and encouragement The various values that I tried to learn from him shall remain a source of inspiration for me forever I owe it to my teachers, particularly Professors R Sharan, PR K Rao, VP Sinha, M U Siddiqui, G C Ray, R K Bansal, Mrs S Gupta, G Sharma and A Mahanta for the deep insights given through the various courses they taught, their advice, encouragement and interest in our work through the years

Professors J K Bhattacharjee and K Banerjee of the Physics department have regularly offered a course on Chaos Theory since the early stages of development of the subject itself I had credited the course during my Undergraduate study Their foresight and lucid exposition of the subject helped me to embark on this topic of research

I would like to acknowledge the help of many persons during the progress of the research work First and foremost, I would like to thank Dr Rakesh Mullick who saw to it that our research work was never delayed by promptly mailing numerous reference papers and other study material whenever I needed them I would also like to thank Dr P V Dhamija of the Central Institute of English and Foreign Languages, Hyderabad and Dr. K. Sriram and his colleagues at Bell Laboratories, U S A, for providing the two speech databases used throughout this work, Dr S Prasad of I I T. Delhi for providing CELP decoded speech data and Dr Preeti Rao for helping in evaluating the quality of decoded speech and making suggestions for the codec studied by us

I have benefitted greatly from the interactions with many research scholars and the cooperation extended by them whenever required For this, I would like to thank Mr Ajit Chaturvedi, Abhay Karandikar, Venkatesh, Deepak Murthy, Chaitanya Babu, Jitendra Das, Sudipto Mukhopadhyaya, Puranjoy Bhattacharya, Balvinder Singh and

Apu Sivadas I will specially cherish my association with fellow research scholar Mr Samarendra Dandapat who has been a constant companion in this endeavour

Finally, I thank everyone who made this work possible and the experience enjoyable

Synopsis

Speech signal coding has been has been an active field of research for over a couple of decades and continues to be so in spite of the increasing proliferation of optical transmission media of relatively unlimited bandwidth This is because of the continued and in fact increasing use of bandlimited media such as satellite lines and radio channels and bit limited storage media such as CD-ROMs Also, the applications of speech coding have become numerous in recent times A major effort has been given in the last ten years to the development of a class of analysis – by – synthesis coding schemes for low and medium bit rate speech coding Most medium and low bit rate speech coders are based on the speech production model of a time varying linear filter excited by a source Such coders are usually designed to estimate the linear filter coefficients and the excitation sequence in a frame – by – frame manner such that the output of the filter approximates the speech signal in some sense. A large portion of the recent research effort has been directed to the design of appropriate excitation functions rather than to investigate alternatives to the linear filter form However, the above paradigm for speech coding may now be approaching a stage of saturation as far as improvements in terms of performance parameters are concerned Further gains in speech coding are likely to accrue by incorporating deeper physiological aspects of the human speech production mechanism and characteristics of the speech signal in the coder structures Towards this end, we have done a dynamical analysis study of speech signals and explored some nonlinear representational forms for predictive coding of speech

The thesis documents our investigation of a *nonlinear* framework for speech signal coding The complete study can be classified as an investigation of three related problems The first problem is to choose a sufficiently general framework for nonlinear speech processing Specifically, we use a deterministic state space framework The choice of a deterministic framework rather than a stochastic one is because of our interest in modelling the time waveform itself instead of its statistical

moments The motivation for using a deterministic state space framework is due to the recent advances in the understanding and characterization of the complex behaviour of deterministic chaos in dynamical systems Viewing complex time series behaviour as arising out of low dimensional chaos gives a new tool for analysing and modelling it deterministically In this framework, the speech time series is embedded in a *reconstructed* state space as a *reconstructed trajectory* We have done a detailed analysis of the reconstructed trajectories of unit articulations of speech, namely phonemes, in terms of dynamical attributes such as dimension, metric entropy and Lyapunov exponents Just as a correlation analysis helps in a linear modelling exercise, these dynamical attributes help in building nonlinear, deterministic state space models As the second problem, we study and compare with linear prediction, the performance of some *adhoc* nonlinear state space based predictive models for speech We have also implemented and carried out preliminary performance tests of a *local state prediction* based low to medium delay speech coding scheme in the medium bit rate range The third problem addresses a related question of estimating the minimum rate at which information about a source can be transmitted to the user subject to the condition that it can be reproduced with a specified distortion We give an algorithm to compute a lower bound of the rate distortion function for stationary ergodic sources *with memory* Both discrete and continuous alphabet sources are considered. Finally, we use this algorithm to compute the lower bound for quantized speech sources.

In the following, we give a chapterwise summary of the thesis Chapter 1 begins with a contextual review of speech coding Thereafter, we build a case for the study of nonlinear analysis and modelling of speech in terms of (a) observations from the speech production mechanism, (b) observations from the speech signal, (c) limitations of a linear model, and (d) advances in nonlinear analysis and modelling techniques We give a qualitative discussion of the notions of randomness, determinateness and predictability in deterministic dynamics, particularly in the light of chaos theory A brief discussion of the three problems in the thesis is given next These are (i) nonlinear dynamical analysis of speech signals, (ii) state space predictive modelling of speech, and (iii) computation of a lower bound of the rate distortion function for

stationary ergodic sources with memory We also record in a historical note, the recent investigations using tools from nonlinear dynamics for speech signal analysis and studies in nonlinear predictive modelling and coding of speech.

In chapters 2 and 3, we are concerned with nonlinear dynamical analysis study of speech signals Chapter 2 begins with a discussion of the theorems that form the basis for dynamical analysis of time series data These theorems give generic conditions for reconstructing a state space trajectory from a scalar observable of a dynamical system evolution such that the dynamical invariants obtained from the reconstructed trajectory will be the same as those of the original dynamical system We discuss two methods for *optimal* state space reconstruction based on singular value decomposition and redundancy criteria and use them to reconstruct speech trajectories and make observations We also give results of the computation of the *largest* Lyapunov exponent of reconstructed trajectories of phoneme articulations Lyapunov exponents give a coordinate independent measure of the local stability properties of a state space trajectory They asymptotically categorize bounded trajectories into equilibrium points, periodic solutions, quasiperiodic solutions and chaos We compare the largest Lyapunov exponent results for speech with those obtained from synthetically generated periodic and quasiperiodic data From the results and comparison tests, we conclude that reconstructed speech trajectories exhibit exponential divergence on the average

In chapter 3, we give results of the computation of two dynamical invariants, namely the correlation dimension and second order dynamical entropy of speech The notion of dimension in dynamical systems is associated with the number of degrees of freedom that a system possesses The "dimension" attribute of a time series is helpful in a deterministic state space modelling exercise because it gives the necessary and sufficient number of independent state space variables needed to model the data A large dimensionality means that the trajectory is "complex" and has numerous degrees of freedom in which case a random process model may be a better choice A study of various phoneme categories shows that speech is largely a *low* dimensional signal We have also computed the correlation dimension from

a simplified statistical model of a particular vowel utterance The dimension results are qualified with a study of the various sources of error affecting the estimates

The second dynamical invariant in which we are interested in chapter 3 is the metric entropy which quantifies the rate of loss of information about the initial state of a dynamical system as it evolves in time Its relevance in the context of state space modelling is because it is inversely proportional to the average time duration for which a dynamical system (or a time series model) can be predicted from a given initial condition. We have computed the second order dynamical entropy of speech which is a lower bound of the metric entropy The positive values of the second order entropy and the largest Lyapunov exponent (chapter 2) for phoneme articulations both give evidence of the average divergence of nearby speech trajectories of speech in the reconstructed state space

Based on the dynamical analysis results, we have investigated some nonlinear prediction schemes for speech signal modelling and coding in a state space framework in chapters 4 and 5 Chapter 4 begins with a review of the salient features of the analysis – by – synthesis class of linear prediction coders and in particular the CELP coding scheme We give some model based analysis results which make a case for nonlinear modelling of speech Thereafter, we study the performance of some nonlinear representation forms for predictive modelling of speech in a state space framework There are two basic schemes in this framework These are the global and local prediction schemes. In a global prediction scheme which we study in this chapter, the function parameters are optimized over the *entire* state space As a first choice, we investigate the (quadratic) polynomial representation form The basis of comparison with short term Linear Prediction (LP) in terms of segmental prediction gain, is the number of coefficients in the two predictor models We have principally considered two ordering schemes for selection of model terms of the quadratic predictor In the first method, we exhaust all possible terms upto a certain time lag before considering terms which include signal dependence for greater lags The second method is based on orthogonal term selection from a set of candidate terms While the first method does not perform as well as a short term LP in terms of segmental prediction gain, the second method gives a modest improvement over LP

for the same number of model coefficients. In another study on quadratic predictors reported recently, the basis of comparison with LP is the time delay upto which signal correlations are considered rather than the number of model coefficients In this case, the performance of the former is significantly better in terms of segmental prediction gain

In chapter 5, we investigate a Local State Prediction (LSP) scheme for speech In this scheme, the representation form is optimized over a local volume in state space where the prediction is to be done The LSP scheme is studied in terms of segmental prediction gain, plots of the prediction error sequence, their spectrum and autocorrelation function and is compared with the error sequence resulting from (i) short term LP, and (ii) short term plus long term LP In LSP, an appropriately chosen neighbourhood of a "target" point in the reconstructed space will contain trajectory points that are close to it in time as well as those which are approximately an integral number of pitch or formant periods away Thus, a LSP attempts to simulate the functions of both short term and long term linear prediction simultaneously Note that in the LSP scheme studied by us, the local neighbourhood is chosen from an analysis frame length of *previous* data values. The performance of a *local linear* prediction scheme in the above terms can be broadly categorized as lying between short term and short term plus long term LP (where both the predictions are done in forward adaptive mode)

We have done a preliminary study of a framework for low to medium delay speech coding in the medium bit rate range based on LSP It is an analysis – by – synthesis coder operationally similar to CELP and tentatively named as a Vector Excited Local State Prediction (VELSP) coder. The following points highlight the coding scheme and bring out the differences with CELP

(i) LSP is performed instead of LP The LSP is performed using previous *reproduced* speech which is available to the decoder as well

(ii) A single excitation codebook designed from empirical data is used instead of two separate codebooks as in CELP, taking advantage of the prediction property of LSP

(iii) A LSP based coder is naturally suited to low delay coding

(iv) Since a LSP based coder is *nonlinear*, we give a method to incorporate the gain factor

We have implemented and studied a basic form of the coder structure at rates of 5.2, 6.5 and 8.0 Kbits/s having theoretical coding delay of 7.5ms, 6.0ms and 4.875ms respectively. While the segmental SNR performance is similar to CELP, the reproduced speech has perceptible noise and becomes poor at the 5.2 Kbits/s rate. The preliminary investigation suggests that the VELSP coding scheme can be a candidate for a more detailed study in terms of (a) improving the quality of reproduced speech by incorporating perceptual distortion measure, adaptive postfiltering depending on a detailed study of the nature of reproduction error etc., and (b) reducing the computational complexity of the coder and decoder.

In chapter 6, we give an application of the computation of a lower bound, $R^L(D)$, of the rate distortion function, $R(D)$, for stationary ergodic sources with memory. The relevance of the rate distortion function $R(D)$ in the context of signal compression is that it gives the minimum rate R at which information about a source can be transmitted subject to the constraint that it can be reproduced with an average distortion D. Specifically, $R^L(D) = R_1(D) + H - H(X)$ for discrete alphabet sources and $R^L(D) = R_1(D) + h - h(X)$ for continuous alphabet sources, where $R_1(D)$ is the first order rate distortion function, $H(X)\{h(X)\}$ is the entropy {differential entropy} of the source X based on the marginal probability density of the source X and $H\{h\}$ is the entropy *rate* {differential entropy *rate*} of the source X. The estimation of the rates $H\{h\}$ becomes difficult as statistical dependencies for larger time frames are successively considered. We give procedures to estimate the second order entropy rate H_2 ($H_2 \leq H$) and the second order differential entropy rate h_2 ($h_2 \leq h$) using a method of generalized correlation sum which is conjectured to give better estimates than the histogram technique. The procedures are based on extensions of a method to estimate the metric entropy that has become standard in dynamical systems literature in the last ten years. We give examples to show the efficacy of this estimation scheme. We also compute the lower bound of $R(D)$ wit

respect to the mean square error distortion criterion for quantized speech sources of resolution 6, 8 and 10 bits/sample

Finally, we conclude in chapter 7 by summarizing the contributions of the thesis and noting some directions for further investigations in these areas

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis documents our investigation of a *nonlinear* framework for speech signal compression The complete study can be classified as an investigation of three related problems The first problem is to choose a sufficiently general framework for nonlinear speech processing Specifically, we use a deterministic state space framework The speech time series is embedded in a *reconstructed* state space as a *reconstructed trajectory* We have done a detailed analysis of the reconstructed trajectories of unit speech utterances, namely phonemes, in terms of dynamical attributes such as dimension, metric entropy and Lyapunov exponents Just as a correlation analysis helps in a linear modelling exercise, these dynamical attributes help in building nonlinear, deterministic state space models As the second problem, we study and compare with linear prediction, the performance of some *adhoc* nonlinear state space models for speech We have also proposed and carried out preliminary performance tests of a local state prediction based low to medium delay speech coding scheme in the 4 8–8 kb/s range The third problem addresses a related question of estimating the minimum rate at which information about a source can be transmitted to the user subject to the condition that it can be reproduced with a specified average distortion We give an algorithm for the computation of a lower bound of the rate distortion function for stationary ergodic sources *with memory* Both discrete and continuous alphabet sources are considered Finally, we use this algorithm to compute the lower bound for quantized speech sources

1.1 A Brief Contextual Review of Speech Coding

Speech signal compression has been an active field of research for over a couple of decades and continues to be so A glance at the recent proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) and other relevant journals would immediately testify to this fact The primary utility of speech compression algorithms lies in speech source coding for distance communication and signal storage Figure 1 1 shows a general block diagram representation of a communication system that subsumes these two functions of a speech compression algorithm The function of the source encoder is to minimize the necessary bit rate for faithfully reproducing the source signal This is done by removing redundancy in the signal and thereby compressing it The channel encoder seeks to introduce redundancy to the source encoder output for the purpose of error protection The

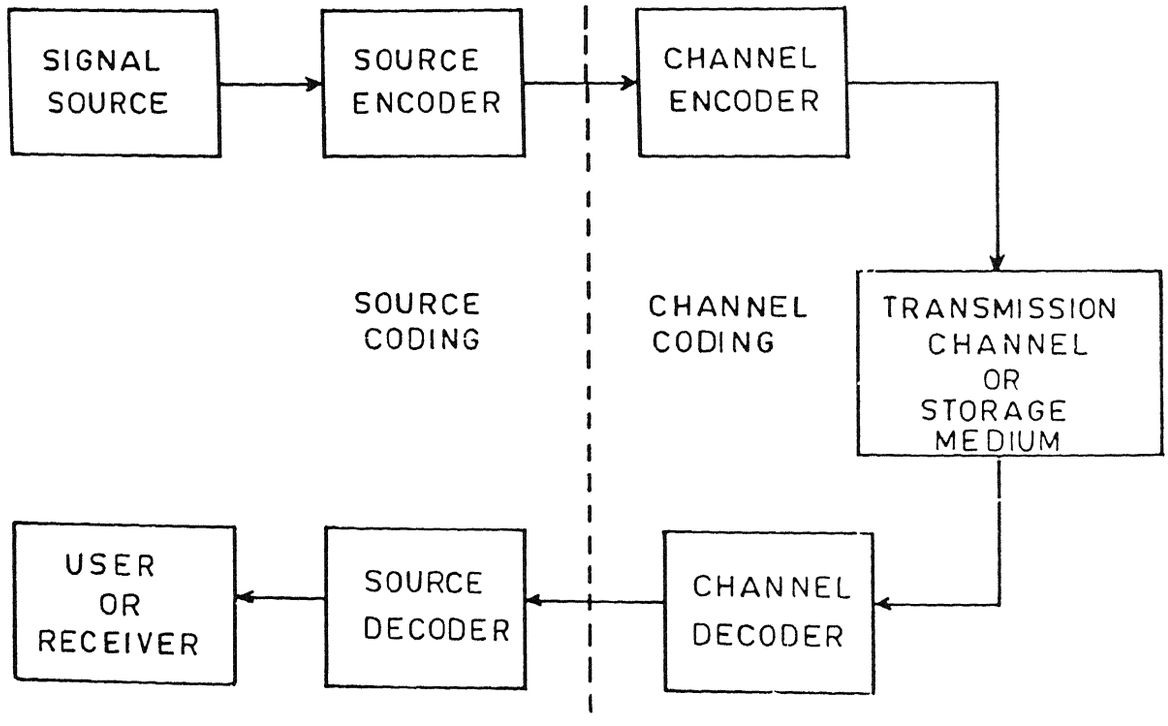


Fig. 1.1: Block diagram of a communication system

channel and source decoders perform the inverse operations of the respective encoders It is often difficult to establish a firm line of demarcation between the successive blocks in the diagram In a speech communication system, the barrier between the source and the source encoder can be at the terminal of a microphone into which a person is speaking in which case it is an analog electrical signal, but more often in a digital environment, the input to the source encoder is considered to be a prefiltered and finely quantized version of the analog signal. Furthermore, to increase the overall efficiency of a communication system, the functions of the source and channel encoder are sometimes integrated together.

Signal compression has remained a frontier research area in spite of the increasing proliferation of optical transmission media of relatively unlimited bandwidth This is because of the continued and, in fact, increasing use of bandlimited media such as satellite lines and radio channels and bit limited storage media such as CD-ROMs Also, the applications of speech coding have become numerous in recent times To quote a spoof ([52], pp 11), "Resources such as bandwidth obey a corollary to Parkinson's Law Resource use will expand to meet the resources available" Some applications to benefit from efficient speech coding algorithms are mobile satellite communications, cellular radio, audio for videophones and videoconferencing, universal cordless telephones, interactive PC software, voice message broadcasting etc

Given the proliferation of speech compression algorithms, and even otherwise, it is necessary to judge their performance and compare them with existing algorithms This is broadly done in terms of four parameters bit rate, decoded signal quality, complexity of implementation and communication delay [74]

(1) **Bit Rate:** This is usually measured in bits/sample or bits/second which is the product of the sampling rate and the number of bits/sample The sampling rate is usually slightly higher than twice the signal bandwidth We are interested in the telephony grade of audio bandwidth where the frequency band is restricted from 200 Hz to 3400 Hz thus requiring a bandwidth of 3 2 kHz, and the sampling rate is 8 kHz

(2) **Decoded Signal Quality:** A lossy compression scheme is most likely to cause a degradation in the reconstructed or decoded signal in terms of perceived quality In the case of speech, the perceived quality is usually measured on a five point subjective scale known as the *mean opinion score* or *mos* scale [75], [79], [80] The five points of quality are associated with a set of standardized descriptions bad, poor, fair, good and excellent An average over many speakers, listeners and speech signal segments evaluates the quality.

(3) **Complexity:** It is mainly the computational effort required to implement the encoding and decoding processes and is usually measured in terms of arithmetic operations and memory requirement Other aspects of coding complexity are the physical size of the encoder and decoder, their cost and power consumption

(4) **Communication Delay:** This refers to the total delay for one - way communication i e coding plus decoding delay Low bit rate coding is typically associated with increased complexity and processing delays in the encoder and decoder While communication delay is largely irrelevant for applications such as broadcasting and storage and message forwarding, it is an important constraint in other applications like network telephony where the delay requirement can be as low as an order of few milliseconds

Apart from these primary parameters of performance, there exist other practical considerations before a signal compression algorithm can be considered for implementation in a source coding scheme One such consideration is the degradation in performance of an encoder decoder pair in the presence of typical channel noise conditions under which it is expected to perform

Overall speech coding systems are often referred to as toll quality, network quality, vocoder quality etc *Toll quality* refers to high quality reproduction, low delay coders and is usually associated with high rate coders *Network quality* is used to refer to those coders which apart from being toll quality, are capable of performing additional functions such as multiple stages of encoding and decoding of speech and high accuracy transmission of non - speech voice band signals such as modem waveforms and network signalling tones *Vocoder quality* coders are those which

maintain high level of intelligibility in the reproduced speech, but speech quality, naturalness and speaker recognizability are all severely compromised

The development of speech compression technology has been supported with the promulgation of coding standards The premier regulatory body in charge of developing speech coding standards is the Telecommunication Standardization Sector of the International Telecommunications Union, referred to as the (ITU–T) which is the successor of the International Telegraph and Telephone Consultative Committee (CCITT) For a discussion on the methodology of standards development followed by the CCITT, see for example [9], [71] The traditional bit rate for network quality telephony was 64 kb/s This was standardized as the μ-law and A-law Pulse Code Modulation (PCM) [75], [114], in the late 1960's and amended in 1972 Subsequently, CCITT standardized a 32 kb/s Adaptive Differential Pulse Code Modulation (ADPCM) [75], [114], scheme in 1984 and revised it in 1986 [11] An 'almost' network quality telephony grade encoding scheme at 16 kb/s was standardized by the CCITT in 1992 This speech coder is based on a Low Delay – Code Excited Linear Prediction (LD-CELP) scheme [29] At the other end of the coding standards' spectrum is the U S Government standard 2 4 kb/s LPC-10 vocoder [145]. This is mainly used in government and defence communications which require digital encryption over a wide range of transmission media Between the two limits of 2 4 kb/s and 16 kb/s are a wide variety of local standards due to various organizations Some examples are the U S Government Federal standard (FS1016) of 4 8 kb/s CELP based coder [23], the Japanese digital cellular radio standard of 6 7 kb/s based on Vector Sum Excited LP (VSELP) [53], the North American digital cellular radio standard (IS54) of 8 kb/s VSELP and the Pan – European digital cellular radio standard (GSM) of 13 kb/s based on Regular Pulse Excitation with Long Term Predictor (RPE-LTP) coder [146] All these coding schemes are based on analysis – by – synthesis linear prediction coding technique [86], [87] Their performance also lies between vocoder and network quality in that they reproduce high quality speech but involve a fairly large coding delay of 50 to 60 ms

A major effort in the last 10 years has been given to the development of analysis – by – synthesis speech coding for bit rates between 4 8 kb/s and 16 kb/s Coders

designed for this range are usually referred to as "medium bit rate" speech coders [68] Present speech coding research activity is concentrated on the design of speech coders in the 2 4 – 4 8 kb/s range which reproduce natural sounding speech Coders in this range are usually referred to as "low bit rate" coders Design of medium bit rate speech coders with specific properties such as low coding delay is also an active field of research For recent reviews on advances and directions in speech compression research, see [74], [51]

Most medium and low bit rate speech coders are based on the speech production model consisting of a time varying linear filter excited by a source The coders are designed to estimate the linear filter coefficients and the excitation sequence in a frame - by - frame manner such that the output of the filter approximates the speech signal in some sense This model is a gross simplification of the speech production mechanism of the human vocal apparatus (see, for example [38], [42], [114] for a detailed discussion) Let us consider this briefly The human vocal system consists of the *vocal tract* which begins at the opening between the *vocal chords* or *glottis* and ends at the lips The total length of the vocal tract in an average male is about 17 cm The *nasal tract* is connected to the vocal tract by the velum and ends at the nostrils When the velum is lowered, the nasal tract is acoustically coupled to the vocal tract to produce the nasal sounds of speech The sub-glottal system consists of the *lungs, bronchi* and *trachea*, which serve as a source of energy for the production of speech

Speech sounds are classified into three distinct classes according to the mode of excitation. *Voiced* sounds are produced by forcing air through the glottis with the tension of the vocal chords adjusted so that they vibrate in a relaxation oscillation, thereby producing periodic pulses of air which excite the vocal tract This fundamental frequency of vibration of the vocal chords is called the *pitch* frequency Examples of voiced sound include /i/ as in b*ea*d, /a/ as in *a*sk, /w/ as in *w*e etc *Fricatives* or *unvoiced* sounds are generated by forming a constriction at some point in the vocal tract (usually towards the mouth end) and forcing air through the constriction at a high enough velocity to produce turbulence This creates a broad spectrum noise source to excite the vocal tract. Examples include /ʃ/ as in *sh*ip, /s/ as in *s*ap

etc. *Plosive* sounds result from making a complete closure (again, usually towards the front of the vocal tract), building up pressure behind the closure and abruptly releasing it This rapid release of pressure causes a transient excitation Examples of plosives are /t/ as in *t*ie, /p/ as in *p*ie etc. As air travels down the vocal tract, the spectrum of the sound is shaped by the resonance frequencies of the vocal tract tube which are called its *formant frequencies* or simply *formants*.

The speech production model of a source exciting a time varying linear filter is based on this idealized description of the production mechanism The vocal tract (and radiation effects at the lips) are accounted for by the time varying linear filter Its purpose is to model the resonance effects of the vocal tract tube The excitation generator creates a signal that models the glottal wave Linear prediction (LP) analysis [114], [96], [3], [99], is used to determine an all-pole model of the filter which is usually block adaptive. The problem of designing an optimal linear predictor in the mean squared sense is formally equivalent to the problem of estimating the coefficients of an autoregressive model from time series using the minimum mean squared error (m m s.e) criterion and which has a rich literature The all-pole filter model is a reasonably good representation of the vocal tract effects However, nasal sounds and fricatives require both poles and zeros for effective modelling This is taken care of by representing the effect of a zero by including more poles in the transfer function [3]

The major difference between the various coders based on the production model lies in the determination of the excitation signal In vocoders, the excitation signal is either a random noise sequence or a train of pulses whose periodicity is determined by a pitch frequency analysis In the analysis – by – synthesis class of speech coders, the excitation signal is determined in a more complex manner Indeed, the difference in the various coders in this class lies primarily in the method of generation of the excitation signal In Regular Pulse excited Linear Prediction Coding (RPLPC), the pulses in a frame length are equally spaced and their positions are completely specified by the position of the first pulse [87], [88] In Multipulse excited LPC (MPLPC), the locations and amplitudes of a fixed number of pulses in a frame are determined [87], [4] In Code Excited LP (CELP) coding, the excitation sequence

(or vector) is determined from gain and shape codebooks [87], [123] Vector Sum Excited LP (VSELP) coding is a special form of CELP coding, which reduces the search complexity of an excitation vector from a codebook [53]

The time varying linear filter models the short term correlation structure or the spectral envelope of the speech signal The excitation sequence may be explicitly determined as in the above coding schemes or it may be derived from a linear filter which models the long term correlation or the spectral fine structure of the speech signal [86], [87] This filter attempts to exploit the pitch period redundancy in the excitation signal and is particularly effective in the case of voiced speech In the case of CELP coders, the pitch redundancy in the excitation signal is either modelled by an excitation vector from an adaptive codebook [81], [30], or by explicitly using a pitch prediction filter

The above discussion shows that linear prediction is a well entrenched concept in the design of speech coders The structure of these coders based on the idealized speech production model has been a very successful paradigm for speech coding It can be further gauged from the fact that most present day speech coding research efforts are concentrated on fine tuning this model structure Before we develop a framework for nonlinear analysis and modelling of speech for coding applications it is necessary to consider the arguments for such a study in the first place We do so, on a rather qualitative basis, in the following section

1.2 A Case for the Study of Nonlinear Analysis and Modelling of Speech

There are various indications that suggest the time is ripe for bringing in nonlinear tools for analysis and modelling of human speech We consider these in the following points.

(a) Observations from the speech production mechanism

In the idealized speech production model, the time varying linear filter models the vocal tract characteristics and the source models the glottal excitation A detailed model of the vocal tract must consider the time variation of the vocal tract shape, losses due to heat conduction and viscous friction at the vocal tract walls, softness

of the vocal tract walls, radiation of sound at the lips and nasal coupling The time varying nature of the linear filter attempts to incorporate the nonstationary nature of the speech signal. The resonances of the vocal tract are reasonably well modelled by the poles of the linear filter Also it can be shown that the bandwidths of the lowest formant frequencies (the first and the second) depend largely on the vocal tract wall loss and the bandwidths of the higher formants depend primarily on the viscous friction and thermal losses in the vocal tract and the radiation loss [114] However, it will be appreciated that the effects of some of the above factors in the vocal tract are largely nonlinear and a linear filter can hardly be expected to do full justice

There are several nonlinearities involved in the vibration of the vocal folds and the generation of the glottal wave

- Strong restoring forces act at the collision of the vocal folds
- During unvoiced sound utterances, the air flow from the lungs becomes turbulent as it passes through a constriction in the vocal tract [114], [42]
- For many sounds, the vocal tract and glottal source do not interact greatly and changes in the vocal tract configuration do not greatly influence vocal fold vibration However, this is not so in the case of voiced sounds. Experimental observations show that at or near the frequency of the first formant there exists a nonlinear coupling between the source and vocal tract [42] Further, it has been shown that the celebrated two mass model of vocal fold vibrations [73] exhibits bifurcations and chaos [67]
- A relatively recent study compares the performance of the output (model of the glottal waveform) of a linear model with a nonlinear model when excited by a train of pulses [120] It is shown that only a nonlinear model can accurately model the dependence of the output on the amplitude and frequency of the driving function.

(b) Observations from the speech signal

It is well known that the speech signal is not ideally modelled by a Gaussian density function [75], [114] Thus, a linear prediction scheme to estimate the present speech

value as a function of previous values of the speech waveform is not optimal in the mean squared sense Further, there has been no systematic study of the higher order statistics of speech There are various practical problems associated with such an effort However, an experimental investigation shows that the third and fourth order distributions of the speech signal can at best be considered as mixtures of Gaussian distributions where the mixing is controlled by the past of the process rather than a single multivariate Gaussian distribution [48]

Studies of the time waveform or state space plot (Chapter 2) of sustained utterances of vowels show that they are *nearly* periodic, although small perturbations are always present even in successive periods Tracking this small variability is of importance for reproducing the naturalness of speech There are large deviations from periodicity in both normal and pathological voice as well These perturbations are variously termed as hoarseness, harshness, raucous voice, husky or creaky voice etc. and are characterized by the sudden appearance of subharmonics when the fundamental or pitch frequency suddenly drops to one–half or one–third the preceding frequency Such transitions can be regarded as manifestations of bifurcations of the underlying dynamical system The backward transition from a subharmonic regime to a high pitched one (of double the frequency) usually occurs via an episode of nonperiodic oscillations which might possibly be related to a chaotic transient For a detailed discussion of these perturbations, see [67] and references therein

(c) Limitations of a linear model

The observation of speech signal characteristics also suggests the limitations of a linear modelling scheme In speech signal modelling for coding, not all the structure is captured in the model (i e the time varying linear filter) The residual (or the filter excitation) carries information about the fine spectral characteristics of the signal

Given that one is interested in capturing as much *relevant* information as possible in the model itself, it is pertinent to ask the question as to what are the limitations of a linear model? For this, let us consider a time series x_n, $n = 1, 2,$ to be modelled A general finite order model is of the form

$$f(x_n, x_{n-1}, \quad x_{n-d}) = e_n \tag{1 1}$$

where f is some representational function, d is the model order and e_n is the error in model fitting at time instant n. In *stochastic* modelling, the observed time series x_n is considered to be a realization of a random process X_n whose joint probability distribution is possibly known. The best possible model (of the random process) will reduce e_n to a realization of an independent, identically distributed (i.i.d.) process or 'strict' white noise. The aim of stochastic modelling scheme is not to model the time series (or waveform shape) itself, but to model the statistical moments. In a *linear* stochastic model, one can at best capture the second order moments of the time series.

In a *deterministic* modelling scheme, the attempt is to model the waveform shape itself. In this case, the laws governing the time series evolution are supposedly known and e_n (as in eq. (1.1)) represents the variable that encompasses information about all the unknown degrees of freedom at time instant n. An autonomous *linear* dynamical system or difference equation (see, for example, Appendix A and references therein for a review of dynamical systems terminology) is not capable of modelling such interesting dynamical phenomena as limit cycles, jump phenomena, amplitude dependent frequency, chaos etc.

(d) Advances in nonlinear analysis and modelling techniques

There has been a marked increase in research activity in both stochastic and deterministic *nonlinear* time series analysis and modelling since the last decade. This is partly because the theory of linear time series modelling is now fairly well advanced but the ability of linear models to represent patently nonlinear behaviour is limited. Also, a major hinderance in the advancement of nonlinear techniques was the limited availability of computational resources which has been overcome to a large extent in recent years.

In the case of stochastic models, nonlinear extensions include bilinear models, threshold autoregressive models, exponential autoregressive models, state dependent models (SDM) etc. [142], [111], [64]. The estimation of model parameters requires knowledge of higher order joint probability distribution / moments of the underlying process. Consequently, more moment information is sought to be modelled compared to the linear case. Attempts to generalize nonlinear models led to the proposal of

stochastic SDMs [111], [64] which include, as special cases, all the linear (AR, MA, ARMA) models and nonlinear models named above

In applications to speech coding for compression, one is more interested in deterministic modelling A straightforward method to build *nonlinear* deterministic models is to choose an adhoc representational form and use a minimization criterion to fit the given time series to it It would be more desirable to have a systematic procedure for model building For this, we draw upon the theory of dynamical systems The bounded, steady state behaviour of dynamical systems can be categorized into one of the following equilibrium point, periodic solution, quasiperiodic solution and chaos (Appendix A) It is primarily the development of a coherent mathematical basis for the description of chaos in nonlinear deterministic dynamics (for example, see books [63], [34], [125], [103], [117] and review articles [36], [108]) in the last one and a half decades that has facilitated a meaningful attempt of the inverse problem of modelling In this framework, we consider the given time series x_n (in our case, the speech signal) to be a scalar projection (or an *observable*) of an evolving dynamical system (i e the vocal tract system) trajectory The modelling problem is to find a representative time series $\hat{x}_n$ using a phenomenological nonlinear state space model to approximate x_n The efficacy of the model is determined by its predictive capability. Certain observations in the theory of chaotic dynamics can be usefully exploited to build nonlinear state space models We will consider these observations and subsequently the steps to build deterministic nonlinear state space models in the next two sections

Before we end this section on a case for nonlinear modelling of speech, it is imperative that we point out the general shortcomings of a nonlinear modelling scheme as well The following points are worth noting

1. There are infinitely many nonlinear representational forms
2. There does not exist a global theory of nonlinear modelling and filtering The approach is to study specific representational forms and develop the theory for those which give better performance in terms of desirable functions
3. The superposition principle is not applicable to the nonlinear case

4 The stability of a linear model is determined by the transfer function only However, for the nonlinear case, stability depends on the initial conditions and the excitation function as well Unlike a linear model, a nonlinear model may be stable in some region of the state space and may not be so in another region

1.3 Randomness, Determinateness and Predictability in Deterministic Dynamics

Traditionally, the theory of deterministic dynamical systems and the theory of random processes have been treated as separate subjects with little or no overlap It used to be implicitly assumed that "random" behaviour was due to extreme complexity of the underlying system If a system is sufficiently complicated, with a large number of irreducible degrees of freedom, then from a practical point of view it becomes impossible to model deterministically – it would be simply not feasible to make enough measurements, much less simulate the model In this case, a random process model allows us to do the best job we can by lumping many degrees of freedom into probability distributions involving only a few variables. Due to the ignorance of the neglected degrees of freedom, predictability is limited, but the model is at least tractable

The first lesson of chaotic dynamics is that randomness does not necessarily involve an enormous number of independent degrees of freedom In the presence of nonlinearity, only a few independent variables are sufficient to generate chaotic motion A chaotic time series can pass all "linear" tests of randomness The second lesson of chaos is that apparent random behaviour may be deterministic but at the same time determinism does not imply predictability upto infinite time As an illustration, consider the binary left shift map,

$$x_{n+1} = 2x_n \bmod 1, \quad x_0 \in (0,1) \tag{1 2}$$

which is piecewise linear on $(0, 1/2)$ and $(1/2, 1)$. Given the initial condition x_0, this map appears to be guilelessly determinate This simple difference equation has an equally simple analytic solution

$$x_n = 2^n x_0 \bmod 1 \tag{1 3}$$

Let the initial condition $x_0 \in (0,1)$ be specified as a binary representation (If r_0 is rational, then its binary representation is asymptotically periodic However, if r_0 is irrational, its binary representation never repeats itself Moreover, irrational numbers form a full measure set on the unit interval) The forward iterates of eq (1 2) are generated by merely moving the decimal point sequentially to the right, each time dropping the integer part to the left of the decimal point Almost all trajectories of this deterministic equation are chaotic Usually the initial condition is known with finite accuracy, say b bits. In such a case, iterating eq (1 2) causes information about the initial condition to be lost after b iterations Thus, although we have a deterministic evolution law and a specified initial condition, exact prediction after some iterations is not possible due to the finite accuracy of the initial condition Only probabilistic statements can be made about the trajectory after b iterations Coupled with the problem of finite accuracy of initial conditions, it is the *sensitive dependence of trajectories on the initial condition* and *folding of trajectories* [63], [34], [125], [36], [108], [117] the two features which may be found in simple nonlinear dynamical systems, that give rise to complicated, aperiodic solutions.

Thus, chaos is a double edged sword On the one hand it tells us that apparent random behaviour in time series may possibly be modelled in a compact, deterministic dynamical system or difference equation model involving a few degrees of freedom, but on the other hand it cautions us that predictability of the model may be limited in time When only a few degrees of freedom are involved we can model the short term behaviour deterministically. In such cases, one can make short term predictions that are possibly better than those from a random process model Thus, apart from capturing regular time series behaviour (periodic and quasiperiodic), it is the possibility of modelling complicated aperiodic behaviour through nonlinear dynamical models that makes this approach both exciting and relevant for such applications as data compression The reader is directed to an interesting article on the issue of randomness of deterministic systems, titled "How random is a coin toss?" by J. Ford [43] Also relevant in this context are references [37], [41], [83]

With the onset of a systematic study of chaotic behaviour in dynamical systems, several criteria have been developed to determine the degree of complexity or

randomness of time series Let us look at them briefly in the perspective of the various interpretations of the concept of randomness through the ages The main thesis invariably consists of regarding randomness as the "absence of laws" Several viewpoints regarding randomness have now been formulated, most of which are qualitatively similar

The first one is the set theoretical approach on which the modern theory of probability is based In this approach, the concept of randomness is associated with the possibility of ascribing to a given quantity a probability measure, i e a quantity is said to be random if it is determined by its probability distributions The "absence of laws" is reflected here by its spread Determinate quantities correspond to distributions described by δ functions In a second order analysis, uncorrelatedness is often equated with randomness However, an uncorrelated process need not be unpredictable Consider, for example, successive iterations of the nonlinear map $x_{n+1} = 4x_n(1 - x_n)$, $x_0 \in (0,1)$ The resulting sequence is a realization of an ergodic process whose autocorrelation function $< x_n\, x_{n+k} > - < x_n >^2 = 0$, unless $k = 0$ Hence, it is uncorrelated, but at the same time predictable In as much as we are concerned with predictability of models and data compression, the most general notion of randomness in this framework is associated with i i d processes which are completely unpredictable

An alternative approach to this concept is the interpretation of randomness as an algorithmic complexity This approach was developed independently by Kolmogorov, Chaitin and Solomonov [28] As a measure of complexity of a given binary sequence x_n, $n = 1, 2, \ldots, N$, it is proposed that one take the length of the program L (in bits) which generates the sequence x_n If the program is small, then L is significantly shorter than the length of the sequence x_n, $n = 1, 2, \ldots, N$, so that it can be regarded as nonrandom In the opposite case, when $L \sim N$, the program essentially reduces to recording the sequence x_n itself symbol by symbol The complexity of the corresponding program can serve as the basis for classifying a given sequence as random Maximum complexity sequences are so unpredictable and incompressible that the term "random" seems appropriate From the viewpoint of complexity, almost all sequences x_n are random since simple programs form a set of measure zero, much

like rational numbers on the real line There is also the practical problem of actually estimating the complexity of a sequence or time series An additional difficulty is the invariable presence of "noise" which makes even algorithmically simple processes algorithmically complex

In experimental situations various empirical notions of randomness are used Some qualitative ones are *nonreproducibility* (impossibility of obtaining the same realization of a process under identical external conditions), *nonrepeatability* (which can be interpreted both as nonreproducibility and the absence of periodicity in the given process), *noncontrollability* and *nonmonitorability* (impossibility of creating conditions under which the process would proceed in a prescribed manner) On a quantitative level are the time average definitions corresponding to statistical ensemble based descriptions of randomness The most widely used criteria are based on associating the quantification of randomness with the degree of absence of periodicity. Examples of this are the criteria of "decaying correlation" or that of a "flat spectrum" While these criteria can distinguish between periodic and aperiodic behaviour in time series, they are not sufficient to differentiate between aperiodicity (deterministic chaos) and randomness as discussed earlier in this section

The answer to whether a time series is truly random (in the sense of unpredictability) or is aperiodic but chaotic (and hence predictable) is provided by estimating nonlinear dynamical invariants such as dimension (fractal, information, correlation dimension etc), metric entropy and Lyapunov exponents of the underlying system [37], [41], [40], [125], [61]) The notion of dimension in dynamical systems is associated with the number of degrees of freedom that a system possesses The number of state variables needed to describe a dynamical system is known as the *nominal* degrees of freedom While a dynamical system may have many nominal degrees of freedom, the trajectories may settle down on or approach a subset of the state space called the *attractor* Attractors exist only for dissipative dynamical systems and can be any one of four types, namely, fixed point, limit cycle, q-torus for quasiperiodicity and strange attractor in the case of chaotic dynamics The dimension of these objects on which the steady state dynamics settles down gives the *effective* degrees of freedom. A deterministic state space model of the steady state behaviour

of a dynamical system is advantageous over a random process model in terms of predictive capability, data compression etc only if the effective degrees of freedom is small, which in turn ensures that a few variables are needed to model the steady state trajectory A large dimensionality means that the trajectory is "complex" and has numerous degrees of freedom In such a case a random process model may perform better

Metric entropy quantifies the rate of loss of information about the initial condition as the dynamical system evolves The *predictability time* of a dynamical system is proportional to the ratio of the logarithmic precision of the initial condition to the metric entropy Lyapunov exponents give a measure of the local stability properties of a trajectory If a trajectory evolves in n-dimensional state space, then the characterization is done through n exponents, usually arranged in decreasing order They categorize bounded trajectories into equilibrium points, periodic solutions, quasiperiodic solutions and chaotic solution A positive Lyapunov exponent of a bounded trajectory indicates that it is chaotic For one dimensional maps, the metric entropy is equal to the Lyapunov exponent The binary left shift map of eq (1 2) has a metric entropy of one bit per iteration Thus, its predictability time is proportional (equal, in this case) to the logarithmic precision of the initial condition, i e b bits

While the definitions of dynamical invariants refer to the dynamical system in question, their estimation from time series (the *observable* of the dynamical system evolution) is made possible by a set of powerful theorems proposed by Takens in 1980 [133] Effectively, these theorems state that the dynamical invariants estimated from the observable time series by suitably reconstructing a state space trajectory will be the same as those of the underlying dynamical system under certain topological conditions

We classify the problems of state space reconstruction and estimation of the three dynamical invariants, namely, Lyapunov exponents, dimension and metric entropy from unit speech utterances, i e phonemes, as the first problem investigated in the thesis Linear analysis such as correlation and PARCOR function evaluation, modes of vibration from spectral analysis etc are not much relevant in a nonlinear modelling exercise The estimation of dynamical invariants provides answers to

whether speech time series can be modelled by a *simple* deterministic state space model in the first place, and to other related questions

1.4 Deterministic State Space Modelling of Time Series

In this framework, the given time series x_n, $n = 1, \quad , N$ is regarded as a scalar projection (or *observable*) of an evolving dynamical system trajectory The modelling problem involves two steps

1. **State space reconstruction:** A state vector is an information set which fully describes the system at a fixed time instant n If it is known with complete accuracy and if the system is strictly deterministic, then the state can contain sufficient information to determine the future of the system The goal of state space reconstruction is to use the immediate past of the scalar observable x_n, to reconstruct a d–dimensional current state vector x_n^d

2. **Nonlinear function approximation:** A general state space model can be represented in the form

$$\mathbf{x}_n^d = \mathbf{g}\left(\hat{\mathbf{x}}_{n-1}^d, \mathbf{u}_n^p, n\right) \tag{1 4a}$$

$$x_n = h\left(\mathbf{x}_{n-1}^d, v_n, n\right) \tag{1 4b}$$

where $\mathbf{x}_n^d$ is a d–dimensional reconstructed trajectory given by eq (1 8) below The scalar output x_n is then a generalized projection of $\mathbf{x}_n^d$ The factors determining the closeness of the approximations based on some distortion criterion, are the representational forms of $\mathbf{g}$ and h and the inputs $\mathbf{u}_n^p$ and v_n To model chaotic behaviour, the representational form of $\mathbf{g}$ *must* be nonlinear Various nonlinear representations are available for the approximation problem Some prominent forms are polynomials, rational polynomials, wavelets, neural nets, radial basis functions etc

It is worth noting that a linear prediction coding scheme of the form

$$x_n = \sum_{k=1}^{d} a_k x_{n-k} + u_n \tag{1 5}$$

where x_n is the output and u_n is the input excitation, can be cast in the state space form of eq. (1 4) In particular, for a linear time invariant system, eq (1 4) can be written in the form

$$\mathbf{x}_n^d = A\mathbf{x}_{n-1}^d + b^T u_n \tag{1 6a}$$

$$x_n = c^T \mathbf{x}_n^d + d^T v_n \tag{1 6b}$$

For the case of eq (1 5), A, b, c and d are respectively given by

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ & & & \cdot & \vdots \\ 0 & 0 & 0 & 0 & 1 \\ a_{d-1} & a_{d-2} & a_{d-3} & a_1 & a_0 \end{bmatrix} \tag{1 7a}$$

$$b = [00 \quad 01]^T \tag{1 7b}$$

$$c = [00 \quad 01]^T \tag{1 7c}$$

$$d = 0 \tag{1 7d}$$

and $\mathbf{x}_n^d$ is given by

$$\mathbf{x}_n^d = [x_{n-d+1} \; x_{n-d+2}, \; \cdot \; x_n]^T \tag{1 8}$$

Predictive state space modelling can be divided into two categories, namely global and local prediction, according to the method by which the function parameters are estimated In a global prediction scheme, the function parameters are optimized over the vectors $\mathbf{x}_n^d$ over the *entire* state space, whereas in a local prediction scheme, the function g is optimized over a local volume in the state space where the prediction needs to be performed

We have investigated the prediction properties of polynomial global and local linear state space predictive models for speech signals We study the latter modelling scheme in some detail and compare its performance with traditional LPC scheme Based on this study we propose a framework for medium bit rate, low to medium coding delay speech compression algorithm using local state space prediction This constitutes the second investigated problem of the thesis

1.5 Fundamental Limit to Signal Compression

In as much as this is a thesis on a framework for speech signal compression, it is relevant to ask the question as to what is the fundamental limit to the coding rate below which no further meaningful compression is possible *Rate distortion theory* provides a mathematical foundation to this problem of source coding In general, the source – receiver pair (of Fig 1 1, for example) is characterized by a probabilistic model and a fidelity criterion to measure the degradation of the coded output with reference to the source The rate distortion function, $R(D)$, gives the *minimum* rate R at which information about a source can be transmitted subject to the constraint that it can be reproduced with an average distortion D [12], [17] According to the information transmission theorem, it is impossible to obtain an average distortion D or less unless $R(D) < C$, where C is the capacity of the transmission channel For memoryless sources, one can usually compute $R(D)$ analytically if the source probability density function (p d f) is known It can also be computed numerically from source output realizations using Blahut's algorithm [17], [16] For sources with memory, one can capitalize on the inherent statistical dependencies to further reduce the minimum rate needed to achieve a specified average distortion However, the computation of the rate distortion function for sources with memory is a difficult task. Analytically, it is known only for a few joint p d f s such as the joint Gaussian density function and for specific distortion criteria [12]

We consider the computation of a lower bound $R_L(D)$ of the rate distortion function $R(D)$ for stationary ergodic sources with memory Specifically, $R_L(D) = R_1(D) + H - H(X)$ for discrete alphabet sources and $R_L(D) = R_1(D) + h - h(X)$ for continuous alphabet sources, where $R_1(D)$ is the first order rate distortion function, H (h) is the entropy rate (differential entropy rate) and $H(X)$ ($h(X)$) is the entropy (differential entropy) based on the marginal probability density of the source X Even in the computation of this lower bound, the estimation of the rates H (h) from finite data length N using histogram technique becomes difficult as statistical dependencies for larger and larger time frames are successively considered We give procedures to estimate the second order entropy rate H_2 ($H_2 \leq H$) and the second

order differential entropy rate h_2 ($h_2 \leq h$) using a method of generalized correlation sum.

Many p d f models have been suggested for the speech process based on first order histograms The Gamma p d f based on the long term statistics, the Laplacian p d f based on the medium term statistics and the Gaussian p d f. based on the short term statistics are among the more popular ones [75]. An estimation of the first order rate distortion functions based on these p d f s and the absolute value and the m s e. distortion measures are available in [2] We have numerically computed the lower bound of the general rate distortion function (i e for sources with memory) with respect to the m s e distortion criterion for quantized speech sources of resolution 6, 8 and 10 bits/sample using our proposed method sketched above

These investigations constitute the third problem studied in this thesis.

1.6 A Historical Note

In this section, we attempt to record the recent investigations using tools from nonlinear dynamics for speech signal analysis and studies in nonlinear predictive modelling of speech The preliminary proposition that advances in nonlinear dynamics especially in the development of tools for the analysis of chaos can be utilized in speech signal analysis in the context of modelling, coding or compression was made primarily in Tishby [140], Kumar [90], Kumar and Mullick [92], [91], [93], Maragos [98] and Bernhard and Kubin [13] The analysis of speech signal complexity in terms of dimension is reported in Tishby [140], Kumar [90], Kumar and Mullick [92], [91], Townshend [143], [144], and Bernhard and Kubin [14] Togneri *et al* [141] show that the space of trajectories of speech may be approximated by a four dimensional manifold which is nonlinearly embedded both in a space of LPC coefficients and in a filter bank space It has been argued and shown by Bandt and Pompe [7] that entropy profiles of speech signals provide a fuller description of its structure compared to the spectrum This is because entropy profiles are invariant with respect to a large class of *nonlinear* distortions of the signal.

There have also been some efforts toward the study of nonlinear predictive modelling schemes for speech with an eye to coding In the category of global

predictive models, neural network based predictive models for speech have been studied and proposed by Tishby [140] and Thyssen *et al* [139], fractal interpolation models by Maragos [98], polynomial difference equation models by Quatieri and Hofstetter [113], Kumar [90], Kumar and Mullick [92], [91], second order Volterra filter models by Thyssen *et al* [139] and recurrent nets by Wu and Niranjan [150] In the category of local predictive models in state space are the pattern search prediction scheme due to Bogner and Li [18], the codebook prediction schemes of Wang *et al* [148], and Singer *et al* [129], the local prediction model of Townshend [144], the Compromised Overlapping Neighbourhood Local Approximation technique of Kumar [143] and Kumar and Mullick [92], [93] and the nonlinear oscillator model of Kubin and Kleijn [89]

Most of the above investigations in nonlinear prediction schemes for speech record an SNR improvement of 2 to 3 dB over a comparable linear prediction scheme Moreover, the prediction residual is reported to be significantly "whiter" than the short term LPC residual Some of the local prediction schemes have been incorporated in DPCM and ADPCM speech coders and their performance studied [143], [89], [51] However, a systematic study of nonlinear predictive speech coding in the medium to low bit rate range has not been reported so far to the best of our knowledge.

We will refer to the above papers on dynamical analysis and nonlinear predictive modelling of speech in more detail at appropriate places in the thesis

1.7 Organization of the Thesis

The organization of the three problems investigated in the thesis is as follows In chapters 2 and 3, we discuss about dynamical analysis of speech signals and give results of the estimation of dynamical invariants for speech Some schemes for nonlinear predictive modelling and coding of speech are discussed in chapters 4 and 5 We give an algorithm for numerically computing a lower bound of the rate distortion function for stationary ergodic sources with memory in chapters 6 and derive some conclusions in chapter 7

Chapter 2 begins with a discussion of the theorems that form the basis for dynamical analysis of time series data These theorems give generic conditions for reconstructing a state space trajectory from a scalar observable of a dynamical system We next discuss two methods for *optimal* state space reconstruction based on singular value decomposition and redundancy criteria and use them to reconstruct speech trajectories and make observations. We also study the local stability properties of speech trajectories by estimating the *largest* Lyapunov exponent from the reconstructed trajectories

In chapter 3, we consider the estimation of two dynamical invariants, namely dimension and metric entropy, from time series data. The state space reconstruction theorems of chapter 2 allow us to get an estimate of these invariants from a scalar observable of the time evolution of a dynamical system We discuss the estimation of these invariants using a method of generalized correlation sum The results of the numerical computation of correlation dimension and second order entropy from speech time series are also given and discussed

Chapter 4 is the first of two chapters concerning nonlinear deterministic state space modelling of speech We first review the salient features of the analysis – by – synthesis class of linear prediction coders and discuss the basic structure and analysis steps of a CELP coding scheme Some model based analysis results indicating the presence of nonlinearities in speech are given We also give the analysis steps and results of out experiments on polynomial prediction of speech and its comparison with the usual linear prediction and make some observations

In chapter 5, we study and compare with linear prediction, the properties of a state space based prediction scheme for speech called the Local State Prediction (LSP) scheme A natural method for incorporating LSP in a speech coder is to use it analogous to a backward adaptive scheme We propose a framework for low to medium delay speech coding in the medium bit rate range based on LSP This coder uses an analysis – by – synthesis scheme and is structurally similar to CELP and named as a Vector Excited Local State Prediction (VELSP) coder

In chapter 6, we propose an algorithm for the computation of a lower bound of the general rate distortion function for stationary ergodic sources with memory

Both discrete and continuous alphabet sources are considered. We give examples of the computation of the lower bound from random process realizations. Although speech signal source is time varying, we use the algorithm to get an estimate of the lower bound of the rate distortion function for quantized speech sources with respect to the mean square error distortion criterion.

Finally, we make some concluding remarks in chapter 7 and give some directions for further study.

Chapter 2

Dynamical Analysis of Speech Signals—1

One of the powerful mathematical tools for the analysis of general time series comes from the modern theory of dynamical systems in the form of a set of theorems These theorems give a method for reconstructing the evolution of a dynamical system in state space from the observation of a single degree of freedom We discuss the import of these theorems in section 2.1. While these theorems give generic conditions for reconstructing the state space trajectory from a scalar observable, they leave unanswered the question of how best to estimate the parameters needed for reconstruction We will consider two methods for optimal state space reconstruction and use them for reconstructing speech trajectories in sections 2 2 and 2 3 In section 2 4, we discuss a method for estimating the *largest* Lyapunov exponent from time series and use it to study the local stability properties of speech trajectories

2.1 A Theory of State Space Reconstruction

In this section, we will review the theory that forms the mathematical basis for state space reconstruction from a scalar time series Figure 2 1 gives a schematic description of state space reconstruction for the Rossler system in the chaotic regime The Rossler system is characterized by the set of equations

$$\begin{aligned} \dot{x} &= -z - y \\ \dot{y} &= x + ay \\ \dot{z} &= b + z(x - y) \end{aligned} \tag{2.1}$$

Figure 2.1(a) shows the projection of the trajectory on the (x, y) plane for $a = 0.15, b = 0.2$ and $c = 10.0$. Figure 2.1(b) shows the time evolution of the scalar *observable* which in this case is the x-variable. It is assumed that the scalar observable is a function of some (unknown) variable(s) of the dynamical system. Figure 2.1(c) shows the reconstructed trajectory from the x-variable in the $(x(t), x(t+1.57))$ plane. There is a differentiable equivalence (which is greater than topological equivalence) between the objects on which the trajectories settle in fig. 2.1(a) and fig. 2.1(c). The set of theorems to be presented as Facts below assert that it is possible to *embed* a scalar time series in the reconstructed state space such that the asymptotic properties of the reconstructed dynamics and those of the original dynamical system are the same. The fact that one can get information about a dynamical system by observing the temporal evolution of just one variable has led to the current explosion in research activity in this field. Qualitatively speaking, the observed variable is intimately related to the other dynamical variables and so its time evolution contains information about them. This would then be the basis for estimating dynamical invariants such as Lyapunov exponents, dimensions and metric entropy from block stationary intervals of speech signals. The theory assures us that they would be the same for the underlying dynamical system, i.e., the vocal tract system configuration and the associated excitation during that utterance interval.

We refer the reader to Appendix A and references therein for a brief review of dynamical systems terminology. The problem of state space reconstruction was proposed in [107] and put on a firm mathematical foundation by Takens [133].

Consider a dynamical system defined by a smooth (at least C^2) diffeomorphic map (a C^k bijection, i.e. 1-1 and onto map) $\phi : M \to M$ (if the system is discrete time) or by a vector field $\mathbf{g}$ on M, (if it is continuous time). Here, M denotes a compact manifold on which the system evolves. Let its dimension, $dim(M) = m$,

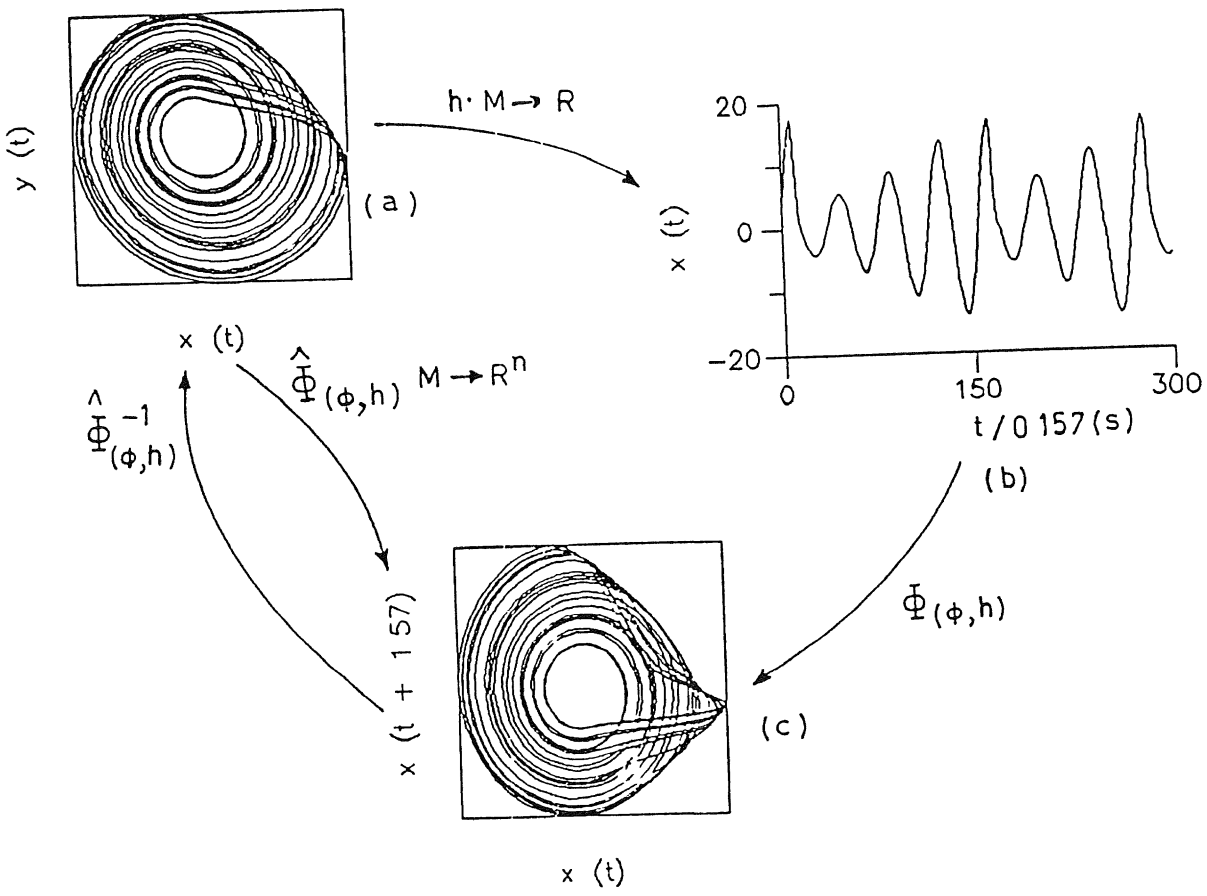


Fig. 2.1: A schematic representation of state space reconstruction using the Rossler system in the chaotic regime (a) Projection of the trajectory on the (x, y) plane, (b) scalar time series corresponding to the time evolution of the x-variable, (c) reconstructed trajectory from the x variable in the $(x(t), x(t + 1\ 57T))$ plane

i e , every point on it has a neighbourhood homeomorphic to R^m The flow of the system is given by $\phi^n(\mathbf{s}_0)$ (discrete time) or $\phi(t, \mathbf{s}_0)$ (continuous time), where $\mathbf{s}_0$ is the initial system state Let $h \quad M \to R$ be a smooth function which induces an *observable* on the dynamical system This produces a scalar time series $r_n = h(\phi^n(\mathbf{s}_0))$ (discrete time observable) or $x(t) = h(\phi(t, \mathbf{s}_0))$ (continuous time observable) Given the above, one can state the following

FACT 2.1: Given a point $\mathbf{p}$ in M, there is a residual (open and dense) subset $C_{\mathbf{g}\,\mathbf{p}}$ of positive real numbers such that for $T \in C_{\mathbf{g},\mathbf{p}}$, the positive limit sets of $\mathbf{p}$ for the flow $\phi(t, \mathbf{s}_0)$ and for the diffeomorphism $\phi(T, \mathbf{s}_0)$ are the same That is, for $T \in C_{\mathbf{g},\mathbf{p}}$, we have that each point $q \in M$ which is the limit of a sequence $\phi(t_i, \mathbf{p})$, $t_i \in R, t_i \to \infty$, is the limit of a sequence $\phi(n_i T, \mathbf{p}), n_i \in N, n_i \to \infty$ ■

FACT 2.2: For a smooth diffeomorphic map $\phi \quad M \to M$, and a smooth function $h \quad M \to R$, it is a generic property that the map $\Phi_{(\phi\, y)} \quad M \to R^{2m+1}$, defined by

$$\begin{aligned}\Phi_{(\phi,y)}(\mathbf{s}_0) &= [h(\mathbf{s}_0)\ h(\phi(\mathbf{s}_0)) \qquad h(\phi^{2m}(\mathbf{s}_0))]^T \\ &= [x_0\ x_1 \qquad x_{2m}]^T \end{aligned} \tag{2.2}$$

is an embedding ■

FACT 2.3: Consider a vector field $\mathbf{g}$, a smooth function h, a point $\mathbf{p}$ in M and a positive real number T For generic $\mathbf{g}$ and h and T satisfying generic conditions depending on $\mathbf{g}$ and $\mathbf{p}$, the positive limit set $L^T(\mathbf{p})$ is diffeomorphic with the set of limit points of the following sequence in R^{2m+1}

$$\Phi_{\mathbf{g},h,\mathbf{p},T} = [h(\phi(kT, \mathbf{p}))\ h(\phi((k+1)T, \mathbf{p})) \qquad h(\phi((k+2m)T, \mathbf{p}))], k = 0, \quad , \infty$$

■

From Fact 2 1 we infer that the limit sets of the continuous time flow $\phi(t, \mathbf{s}_0)$ and those of the corresponding discrete time mapping $\phi^n(\mathbf{s}_0)$ introduced by the process of sampling the flow at uniform intervals T, are the same This is true for most choices of sampling time T, The values of T, that do not preserve the limit sets are those that are commensurate with the system's inherent excitation modes In such cases, a small perturbation of the sampling time removes the problem

Fact 2.2 is of direct relevance to the reconstruction problem. It is based on Whitney's embedding theorem which states that every m-dimensional compact manifold M embeds in R^{2m+1}. *Embedding* means that there exists a diffeomorphism $\Phi \quad M \rightarrow R^{2m+1}$. Here, it is stated that a *specific* map $\Phi_{\phi,h}$, which uses time delays to reconstruct the state space, will work. Takens has also proved that one can use time derivatives to create embeddings. While R^{2m+1} is the largest space needed for reconstructing the dynamics (a sufficiency condition), it is often possible to do so only in R^m (a necessary condition). The space in which the dynamics is reconstructed is called the *embedding* or *reconstruction* space, and its Euclidean dimension is called the *embedding dimension*.

Fact 2.3 follows as a corollary to Facts 2.1 and 2.2. Fact 2.1 asserts that the limit set of a continuous time flow is same as that of a discrete time one obtained by uniformly sampling the flow and Fact 2.2 gives a particular method for embedding discrete time maps on manifolds into R^{2m+1}. Fact 2.3 states that a uniformly sampled data set embedded into R^{2m+1} has the same limit set as that of the original system under observation.

Facts 2.1–2.3 provide a theoretical basis for reconstructing the state space from a scalar observable of a dynamical system. However, in an experimental situation or otherwise, where one is confronted with a time sequence of scalar observations, one does not have *apriori* knowledge of the dimension, m, of the manifold on which the original system dynamics evolves. One way to overcome this problem during state space reconstruction is to increase the embedding dimension systematically until the trajectories in the corresponding embedding space do not seem to intersect. Another practical problem is the choice of an appropriate time delay between the scalar components of the reconstructed vector. Theoretically, one can also choose the successively sampled data values as the constituents of the reconstructed vector as in eq. (2.2), because of the premise that successive scalar measurements contain some new information about the dynamics. However, experimental considerations like noisy measurements etc. require that some optimality condition be used in the choice of time delay. In the next two sections, we review two such optimality criteria

from dynamical systems literature, discuss their relative merits and demerits and use them to plot state space trajectories of speech signals on the reconstructed space

2.2 Optimal State Space Reconstruction using SVD Criterion

The observation process can be represented by

$$\mathbf{s} = \mathbf{g}(\mathbf{s}(t)) + \Lambda(t)$$
$$x(t) = h(\mathbf{s}(t)) + \gamma(t) \tag{2.3}$$

where s is a d-dimensional state vector, h is a scalar valued observation function, $\Lambda(t)$ is a d-dimensional vector which denotes additive *dynamical noise* and $\gamma(t)$ denotes additive *observational noise* $x(t)$ is a scalar valued continuous observable which may be uniformly sampled at intervals of T to produce the observed time series $r(t+iT)$, or x_i, $i = 1, 2,$ Observational noise does not alter the deterministic state of the system whereas dynamical noise does A general time delay reconstruction in R^n produces a reconstructed vector time series

$$\mathbf{x}^n(t+iT) = [x(t+iT)\ x(t+iT+kT) \qquad x(t+iT+(n-1)kT)]^T$$
$$\mathbf{x}_i^n = [x_i x_{i+k} \qquad x_{i+(n-1)k}]^T,\ i = 1, 2, \tag{2.4}$$

Rigorously speaking, the theory of state space reconstruction as discussed in section 2.1 is valid only for infinite amount of noise free data Otherwise, invariants are dependent on the type of reconstruction However, the theory is still useful in the low noise limit case In the following, we will assume that there is no dynamical noise during system evolution, i.e., $\Lambda(t) \equiv 0$, and the only source of randomness is observational noise Observational noise can arise due to finite precision of the measuring instrument, inaccuracy of measurement due to filtering effects etc

The process of embedding the dynamics in a reconstructed state space requires a choice of 3 parameters sampling time T, embedding dimension n and time delay k, as seen from eq (2.4) Qualitatively speaking, the state of a dynamical system contains all the information about it at a particular time instant Therefore, a good state space reconstruction is one in which the state vector contains the

maximum possible information about the dynamics to make its future as predictable as possible It is also desirable that the dimensionality of the state vector be as small as possible. Various state space reconstruction methods have been proposed based on singular value decomposition (SVD) [21], information theoretic criteria [46], [45], other statistical [25], geometrical [27] and topological considerations [94]

The method of singular value decomposition (SVD) was proposed by Broomhead and King [21] for use in state space reconstruction SVD is used to arrive at an appropriate embedding dimension n Heuristic arguments are used to compute the *window length*, $T_w = nkT$, which is the time span of the reconstructed vector (see eq (2 4)) From the observed scalar time series x_i, $i = 1, \quad , N$, we reconstruct l dimensional vectors $\mathbf{x}_i^l, i = 1, \quad , N - l + 1$, where $\mathbf{x}_i^l = [x_i x_{i+1} \quad x_{i+l-1}]^T$, and N is rather large The vector time series is used to form a *trajectory matrix* X

$$\begin{aligned} X &= N^{-\frac{1}{2}} [\mathbf{x}_1^{l\,T}\ \mathbf{x}_2^{l\,T} \quad \mathbf{x}_{N-l+1}^{l\,T}]^T \\ &= N^{-\frac{1}{2}} \begin{bmatrix} x_1 & x_2 & & x_l \\ x_2 & x_3 & & x_{l+1} \\ \vdots & & \cdot & \\ x_{N-l+1} & & & x_N \end{bmatrix} \end{aligned} \tag{2 5}$$

The SVD of the $(N - l + 1) \times l$ matrix $\mathbf{X}$ allows us to decompose it in the form

$$X = U S V^T \tag{2 6}$$

where U is a $(N - l + 1) \times l$ dimensional matrix such that its columns are orthogonal, V is a $l \times l$ orthogonal matrix such that $V^T = V^{-1}$ and S is a $l \times l$ diagonal matrix whose elements $\sigma_1, \quad , \sigma_n$ are nonnegative and are called the *singular values* They are, by convention, arranged in decreasing order, i e , $\sigma_1 \geq \sigma_2 \geq \quad \geq \sigma_l \geq 0$ If rank $(X) = n < l$, then $\sigma_n > 0$, and $\sigma_{n+1} = \quad = \sigma_l = 0$ The columns of the matrix U comprise of the eigenvectors of XX^T, l of which have nonzero eigenvalues The eigenvalues and eigenvectors of the matrix $X^T X$ are the square of the singular values and the columns of the matrix V respectively Also, the $l \times l$ matrix $X^T X$ is the covariance matrix of the scalar components of $\mathbf{x}_i^l$ averaged over the entire trajectory

$$X^T X = \frac{1}{N} \sum_{i=1}^{N-l+1} \mathbf{x}_i^l \mathbf{x}_i^{l\,T} \tag{2 7}$$

If we assume that the elements of the matrix X come from some continuous distribution, then full rank matrices comprise a set of maximal measure [101] Thus, one cannot expect X to be less than full rank The presence of noise in the time series causes a saturation in the singular value spectrum It, for example, a scalar time series x_i consists of a deterministic component $\hat{x}_i$ $(= h(s(t_i))$, see eq (2 3)) and an additive white noise component ξ_i, i e , $x_i = \hat{x}_i + \xi_i, i = 1, \quad , N$, then asymptotically [21]

$$\sigma_i{}^2 = \hat{\sigma}_i^2 + < \xi^2 >, \ i = 1, \quad , l \tag{2 8}$$

where $\sigma_i{}^2$ is an eigenvalue of $X^T X$ and $\hat{\sigma}_i^2$ is an eigenvalue of $\hat{X}^T \hat{X}$, $\hat{X}$ being the trajectory matrix corresponding to the deterministic time series $\hat{x}_i$ The singular spectrum σ_i saturates to a noise floor value when $< \xi^2 >$ is significantly greater than $\hat{\sigma}_i^2$ This defines a threshold on the magnitude of the singular values such that those values less than the threshold correspond to predominantly noise coordinates The embedding dimension n is taken as the effective rank of the trajectory matrix corresponding to the number of singular values above the threshold, since this is the subspace spanned by the deterministic part of the time series

The choice of *window length*, $T_w = nkT$, where n is the embedding dimension, k is the time delay and T is the sampling time, is based on heuristic arguments In order to exclude more than one integer data period within a window length, it is sufficient that, for bandlimited data, $T_w \leq \frac{1}{f_l}$, where f_l is the bandlimiting frequency A lower bound on T_w is given by $T_w \geq (2m+1)T$, where m is the dimension of the manifold on which the dynamics evolves However, since m is not known *apriori*, one can use an estimate $T_w = \frac{1}{f_l}$ This gives an estimate of the product kT which is the time delay, in seconds, between successive scalar constituents of the reconstructed vector in eq (2 4) Usually, the sampling time T is fixed by the measuring instrument, in which case, an integer estimate of the time delay is $k = [T_w/nT]$ where $[\,]$ operation denotes the choice of the largest integer k such that $nkT \leq T_w$

The SVD method was one of the first proposed to reconstruct state vectors from experimental time series using certain optimality criterion However, there are limitations with this method The principal idea here is to determine those

coordinate directions (the singular vectors) along which the embedded trajectory has significant spread (given by the singular values) Recall that the singular values and vectors of X can be computed from an eigenanalysis of the covariance matrix X^TX This means that the orthogonality of the singular vectors translates to *linear independence* or uncorrelatedness in the limit of large *apriori* embedding dimension l Thus, two scalar time series having the same covariance structure but different higher order moments cannot be distinguished by an SVD analysis Moreover if the scalar time series x_i is an observable of a chaotic dynamical system, then the embedding dimension n is equal to the *apriori* embedding dimension l, however large l is This is because in the limit of large l, the coordinates in SVD become the Fourier coefficients and in the case of chaotic dynamics, all the Fourier coefficients are nonvanishing and contain information about the dynamics It is suggested in [101] that the singular values be estimated directly from X rather than from the eigenvalues of the X^TX matrix which may lead to possible numerical artifacts in the form of apparent convergence of the singular values due to machine precision

The efficacy of the SVD criterion in state space reconstruction lies in noise reduction If one knows the noise level or the accuracy of the data *apriori*, then it is prudent to discard that subspace of the trajectory matrix which corresponds to singular values below the noise threshold This method is also useful in plotting 2-d projections of state space plots because ideally it is desirable to project the embedded trajectory onto that subspace in which it has maximal spread

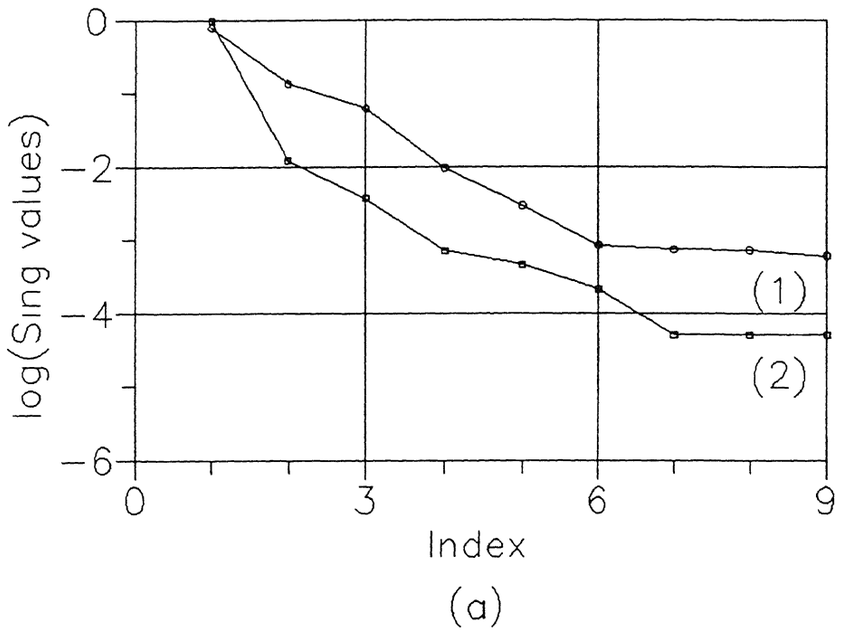
We have studied this reconstruction scheme on unit speech utterances, namely phonemes of database 1 (Appendix B) The SVD analysis was performed on 44 consonants of the International Phonetic Alphabet (IPA) spoken by 4 trained persons (3 males and 1 female) and 8 cardinal vowels spoken 4 times by a single person (Daniel Jones) We excluded the 13 plosives from all analysis reported in the thesis because of their extremely short duration of utterance Thus, a total of 208 ($44 \times 4 + 8 \times 4$) phoneme utterances were used for all analysis work

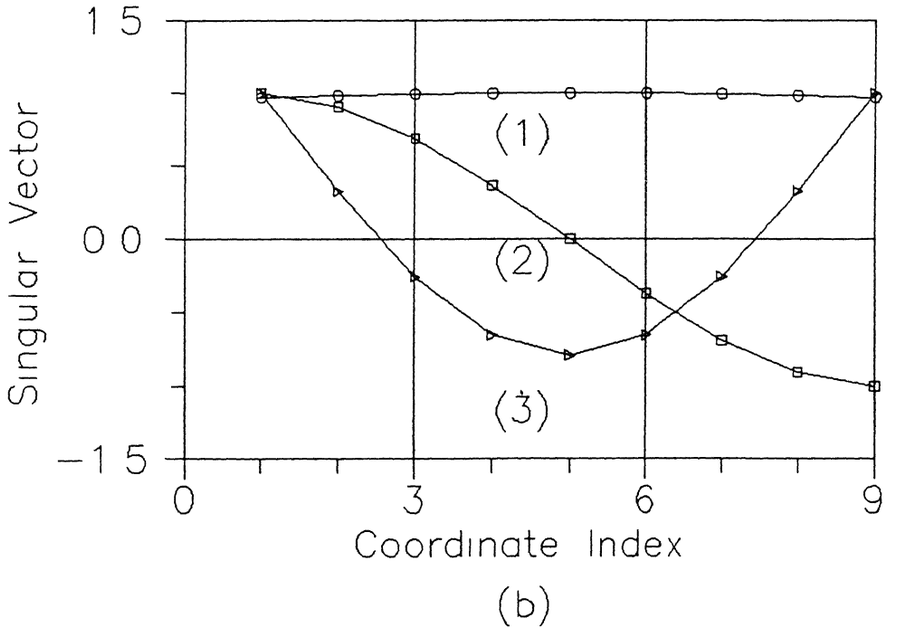
To form the X matrix for SVD analysis, we choose a window length $T_w = 0\,5$ ms which allows us to use an *apriori* embedding dimension $l = 9$ at time delay $k = 1$ Since the highest significant frequency content of speech may be greater than

$1/T_w = 2$ kHz, it is generically possible that the reconstructions are "bad" Choosing T_w according to the argument given earlier will not allow us to retain a sufficiently large l This can be seen as follows – assuming the bandlimiting frequency of speech to be $f_l = 4$ kHz, we will have only 5 samples in the window at 16 kHz sampling rate Thus, the largest allowable value of l is equal to 5 in this case Figure 2 2(a) shows the normalized singular value spectrum in log scale ($\log_{10}(\sigma_i/\sum_i \sigma_i)$, $i = 1, \quad ,9$) of the trajectory matrix X of two time series corresponding to cardinal vowel /a/ and approximant /ɻ/ Figure 2 2(b) shows the first three singular vectors of X corresponding to /a/ These two graphs are typical of the SVD analysis done on all the phonemes of the database The embedding dimension n is chosen to be the number of singular values above the saturation floor and it varies from 3 to 6 No significant difference in the mean value of n is observed across the broad phoneme classes of nasal, trill, tap or flap, fricative, lateral fricative, approximant, lateral approximant and cardinal vowel It is noteworthy that the mean threshold value across the 22 fricatives is a factor of 3 2 higher than that across the 8 cardinal vowels We interpret this as the inability of the SVD scheme to sift the "more complex" deterministic part of the fricative trajectories from the noise component compared to the cardinal vowel case In the following section, we will discuss a reconstruction scheme based on the more general notion of independence rather than that of *linear* independence of the SVD scheme

In figs 2 3–2 7, parts (a)–(d) each, we plot for five phoneme utterances, the respective time series, fourier spectrum and the reconstructed state space trajectories using the SVD and mutual information criteria The five phoneme utterances corresponding to figs 2 3–2 7 are cardinal vowels /ɪ/, /o/ and /a/, fricative /ʒ/ and approximant /ɻ/ respectively In part (a) of each figure, 400 samples of the time series are plotted At 16 kHz sampling rate, this is equivalent to 25 ms of speech utterance In part (b), we plot the first 100 points of the corresponding 400 point fourier spectrum The 2-d projections of the reconstructed trajectory onto the 1–2 plane spanned by the first two singular values are shown in part (c) of figs 2 3–2 7 We will comment on the reconstructed trajectories in part (d) of the respective figures in the next section One can make the following observations from these figures

Fig. 2.





2: (a) The normalized spectrum of singular values for (1) cardinal vowel /a/ and (2) approximant /v/ Data length $N = 4000$ and 1750 respectively Sampling rate $1/T = 16$kHz at 12 bits/sample (b) The first 3 singular vectors corresponding to cardinal vowel /a/

1 For cardinal vowel utterance /i/ (fig 2.3), the pitch period delay is equal to 84 samples. Thus, the projected trajectory of fig 2.3(c) contains slightly less than 5 loops. Figure 2.3(b) shows the presence of only one prominent frequency in the spectrum. This accounts for the approximately circular shape of the reconstructed trajectory in fig 2.3(c). Note that the trajectory never exactly repeats itself. It is felt that the tracking of this minor variability even in periodic and quasiperiodic utterances is important for reproducing the naturalness of speech.

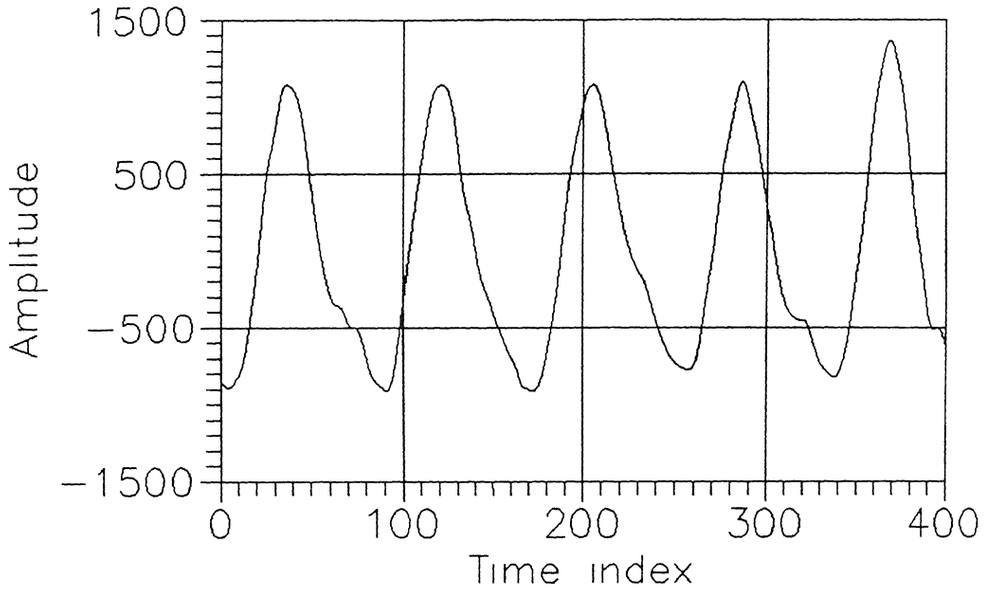
2 For cardinal vowel utterance /o/ (fig 2.4), the pitch period delay is equal to 89 samples. There is a prominent frequency content at approximately double the pitch frequency which causes a looping in each cycle of the trajectory (fig 2.4(c)).

3 For cardinal vowel utterance /a/ (fig 2.5), there are comparatively more number of prominent frequencies in the spectrum. The pitch period delay is equal to 90 samples. There is a prominent frequency at approximately 5 times the pitch frequency. Thus, within each of the four major loops of the projected trajectory (fig 2.5(c)), the three smaller loops are approximately one-fifth the length of the major loop.

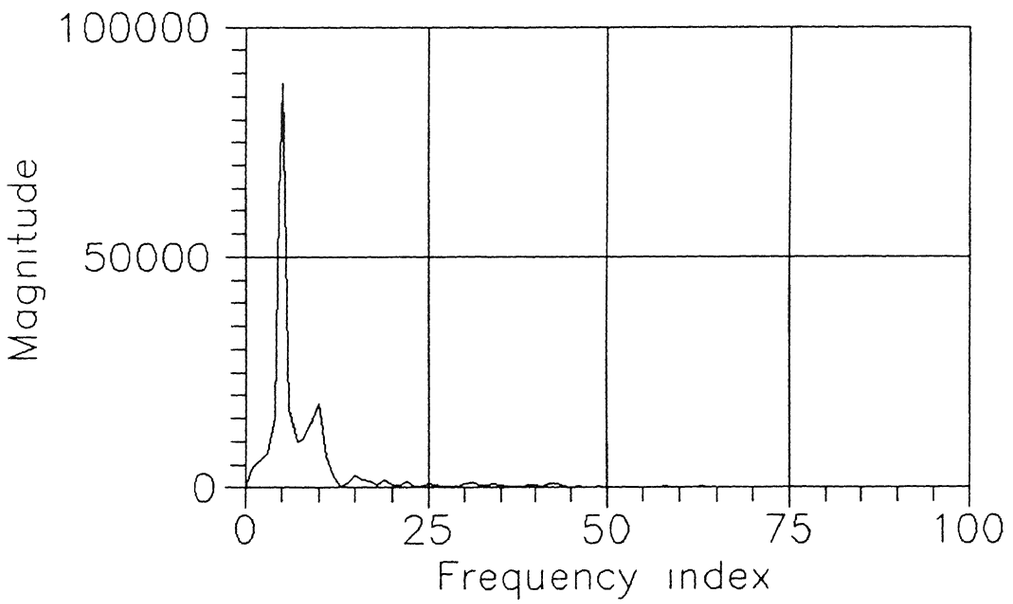
4 For fricative utterance /ʒ/ (fig 2.6), the projection of the reconstructed trajectory (fig 2.6(c)) appears to be "more complex" than that corresponding to cardinal vowels. However, reconstruction using the redundancy criterion (fig 2.6(d)) appears to do a better job than the SVD criterion in this case. We will comment further on this in the next section.

5 For approximant utterance /ɹ/ (fig 2.7), the pitch period delay is equal to 142 samples. At 16 kHz sampling rate, this corresponds to a frequency of 112.7 Hz. The 3 major loops in the projected trajectory each correspond to a pitch period. There are two loops in each major loop corresponding to the prominent frequencies at approximately double the pitch frequency.

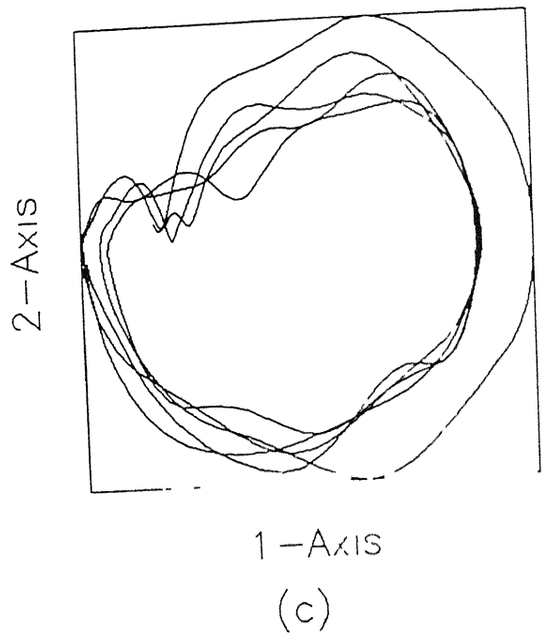
In section 2.3, we discuss another reconstruction scheme that is firmly grounded in theory. We will also make some comparisons between the two reconstruction schemes in the next section.


1500
500
−500
−1500
Amplitude
0
100
200
300
400
Time index

(a)


100000
50000
0
Magnitude
0
25
50
75
100
Frequency index

(b)



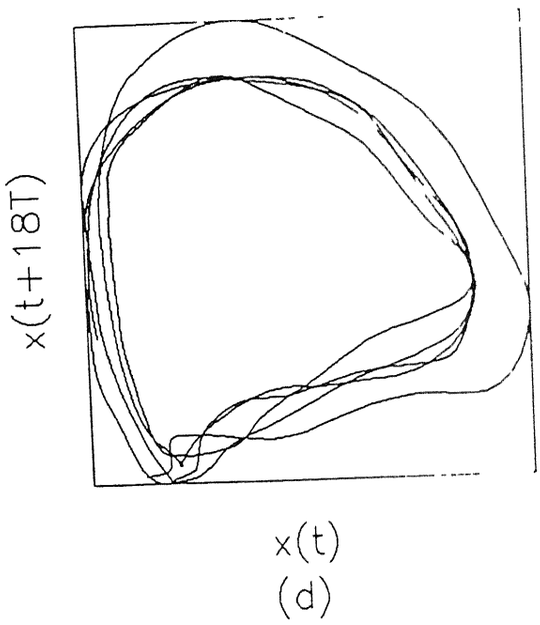
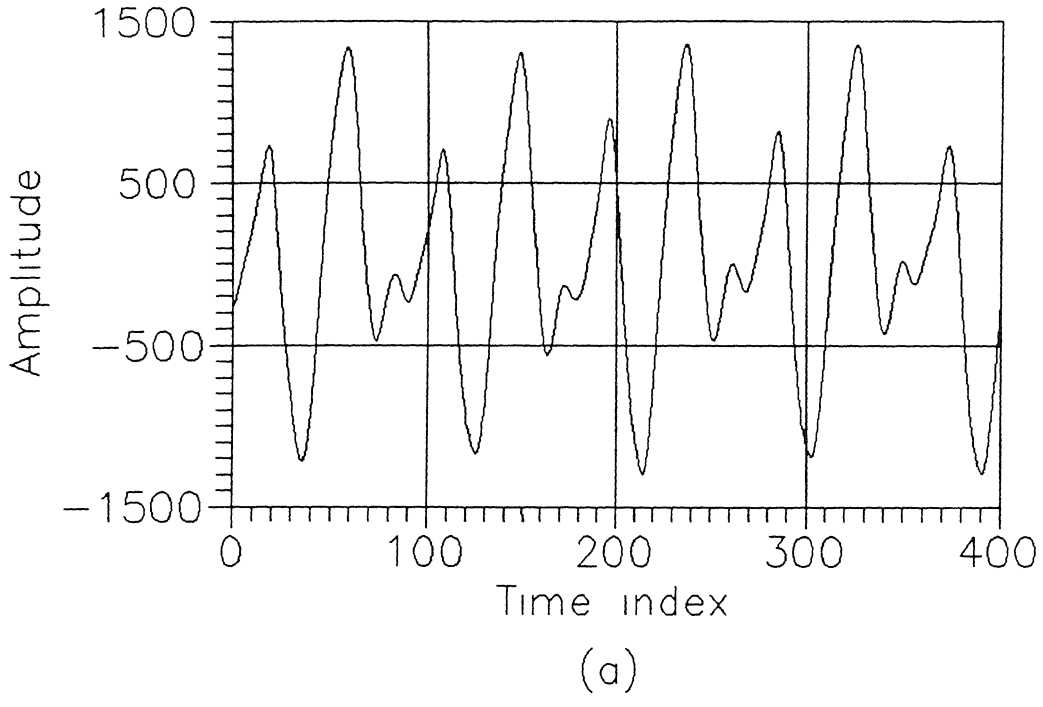
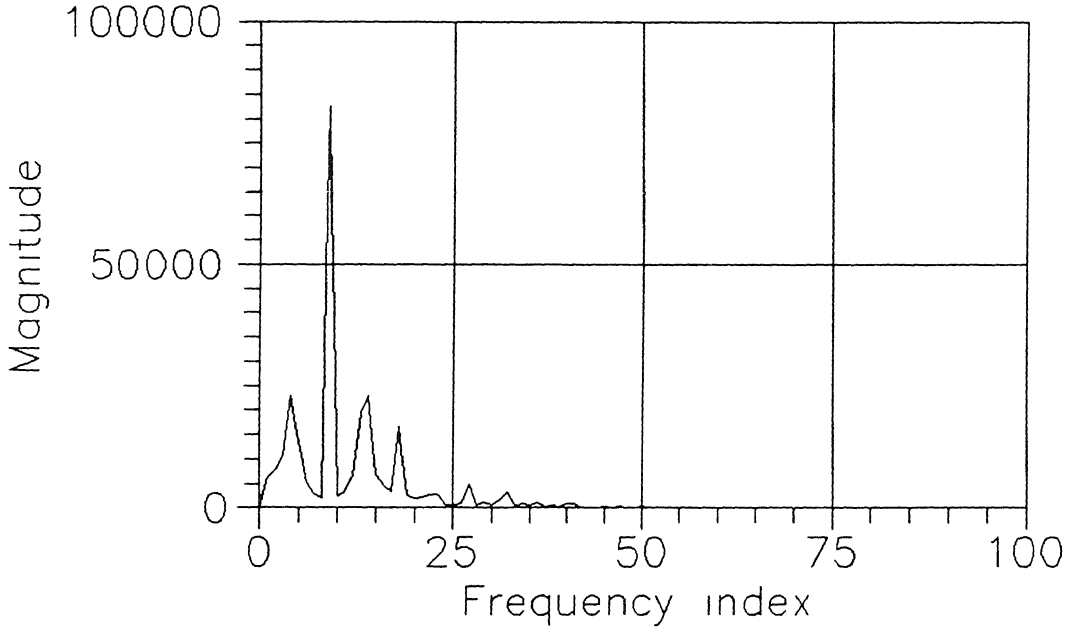


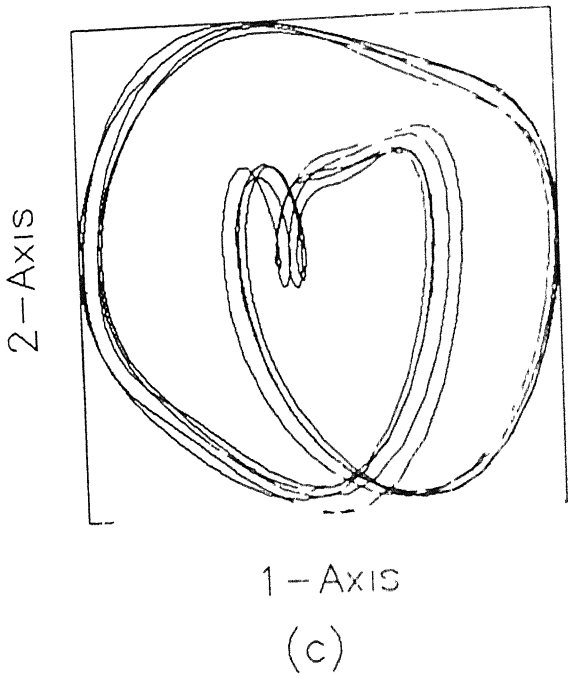
Fig. 2.3: For cardinal vowel utterance /ɪ/, (a) part of the time series at sampling rate $1/T$ = 16kHz, (b) the first 100 points of the corresponding 400 point fourier spectrum, (c) the projection of the reconstructed trajectory using SVD criterion on the 1–2 plane corresponding to the 2 largest singular values, (d) trajectory plot on the $(x(t), x(t+T_d))$ plane where T_d is estimated using minimum mutual information analysis


1500
500
−500
−1500
Amplitude
0
100
200
300
400
Time index

(a)


100000
50000
0
Magnitude
0
25
50
75
100
Frequency index

(b)



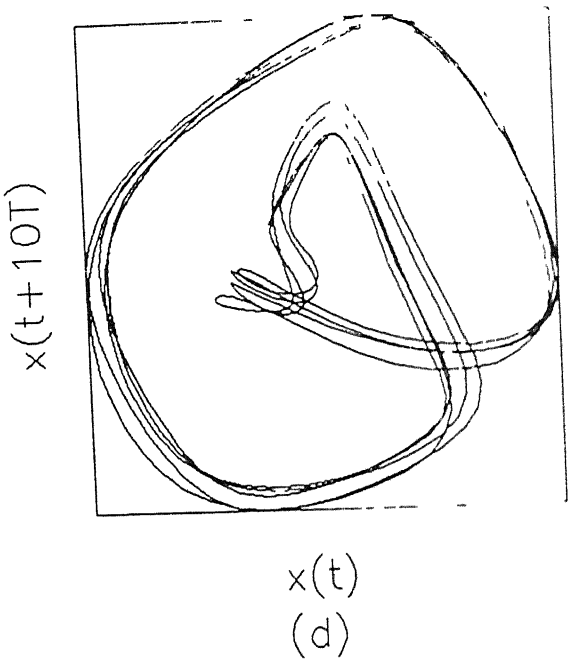
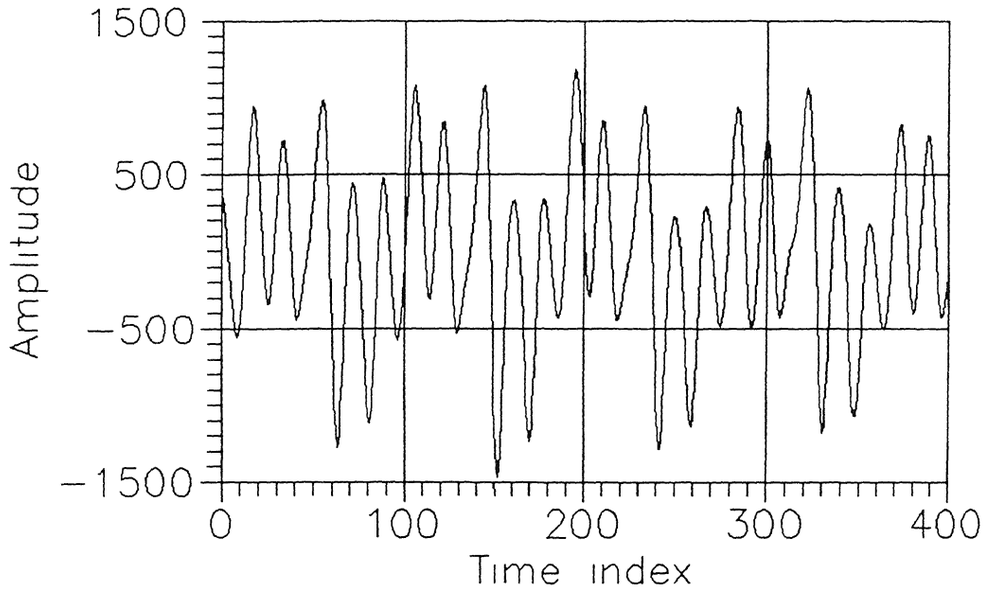
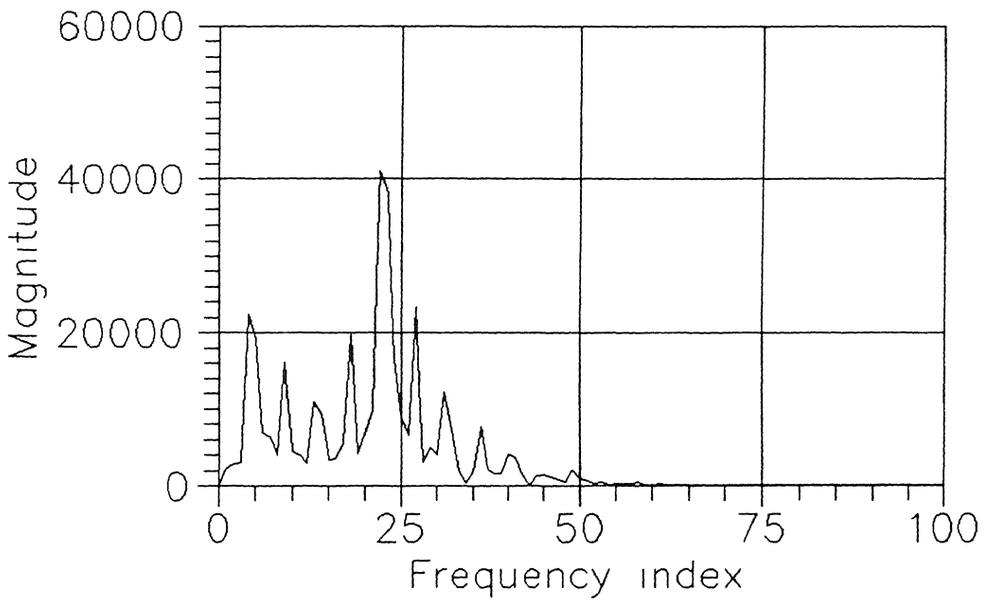


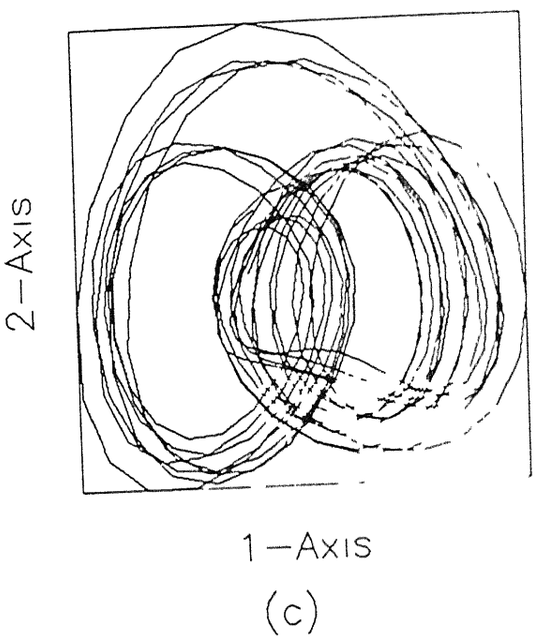
Fig 2.4: For cardinal vowel utterance /o/, (a) part of the time series at sampling rate $1/T$ = 16kHz, (b) the first 100 points of the corresponding 400 point fourier spectrum, (c) the projection of the reconstructed trajectory using SVD criterion on the 1–2 plane corresponding to the 2 largest singular values, (d) trajectory plot on the $(x(t), x(t+T_d))$ plane where T_d is estimated using minimum mutual information analysis


1500
500
−500
−1500
Amplitude
0
100
200
300
400
Time index

(a)


60000
40000
20000
0
Magnitude
0
25
50
75
100
Frequency index

(b)



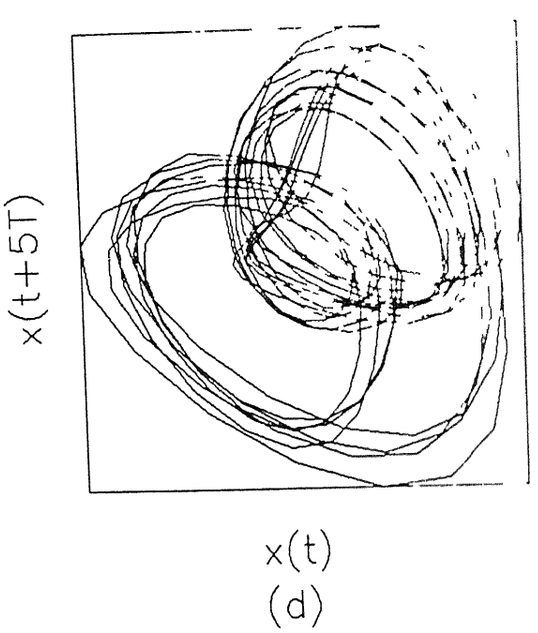
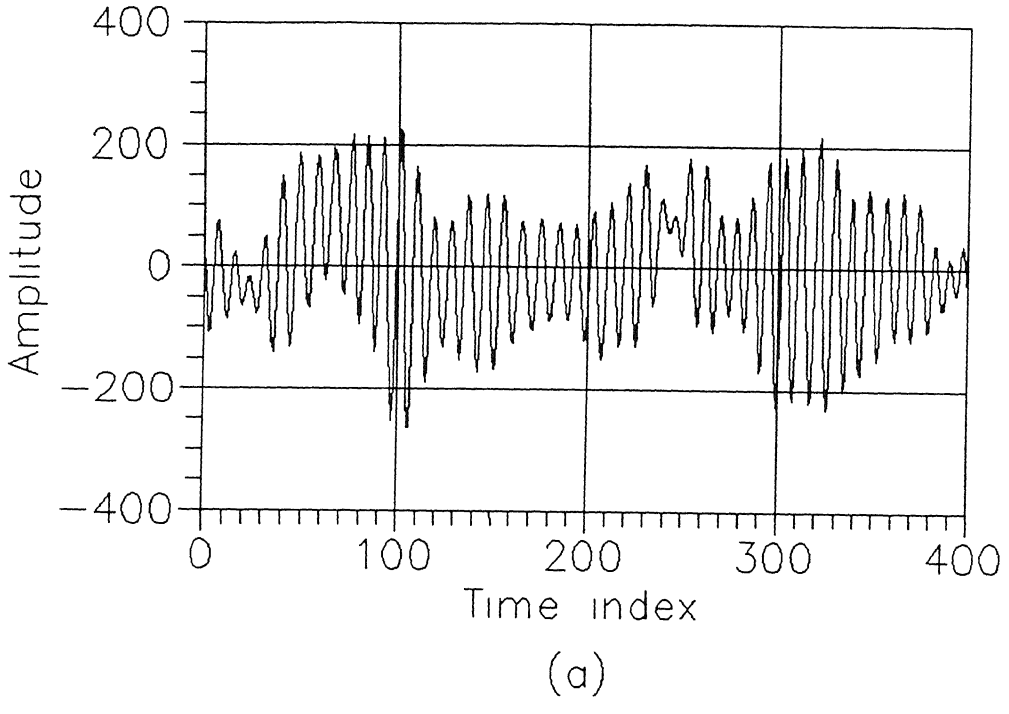
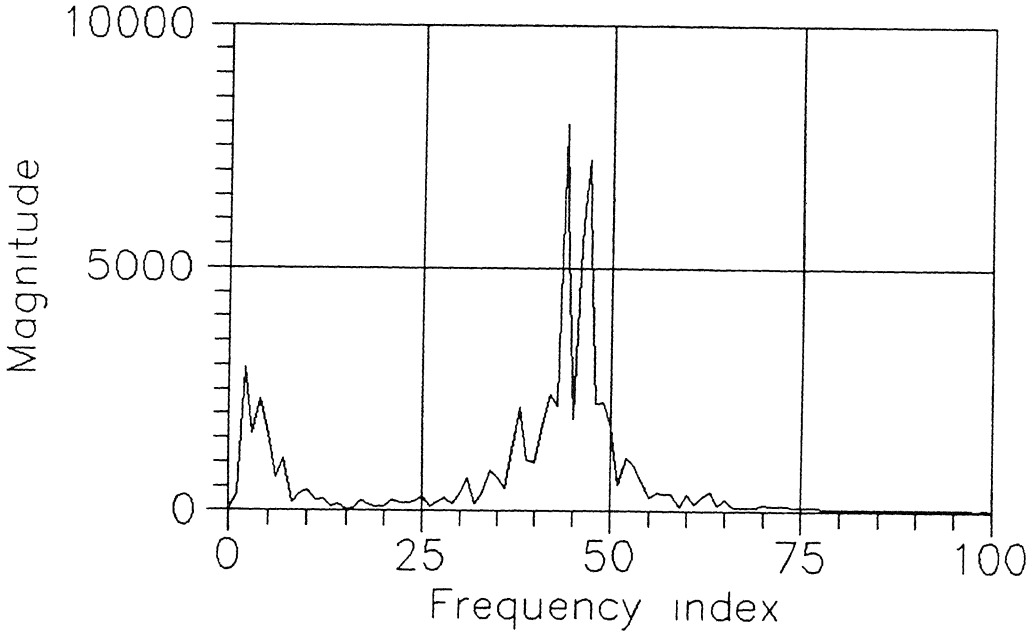


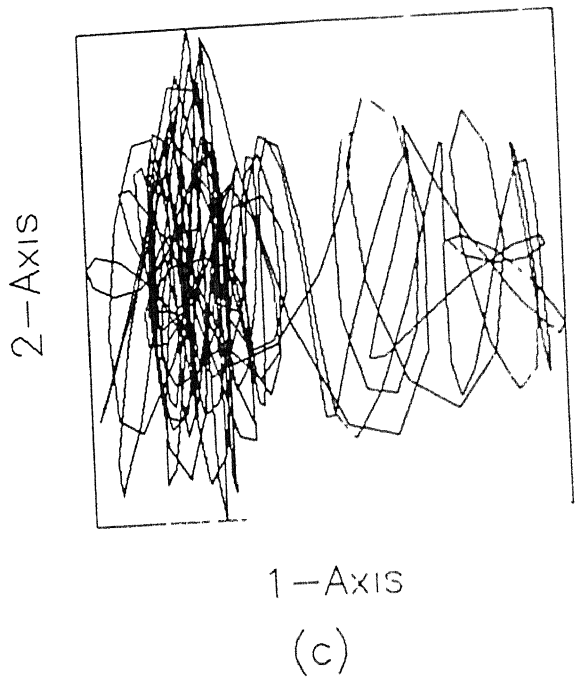
Fig. 2.5: For cardinal vowel utterance /a/, (a) part of the time series at sampling rate $1/T$ = 16kHz, (b) the first 100 points of the corresponding 400 point fourier spectrum, (c) the projection of the reconstructed trajectory using SVD criterion on the 1–2 plane corresponding to the 2 largest singular values, (d) trajectory plot on the $(x(t), x(t+T_d))$ plane where T_d is estimated using minimum mutual information analysis


400
200
0
−200
−400
Amplitude
0
100
200
300
400
Time index

(a)


10000
5000
0
Magnitude
0
25
50
75
100
Frequency index

(b)



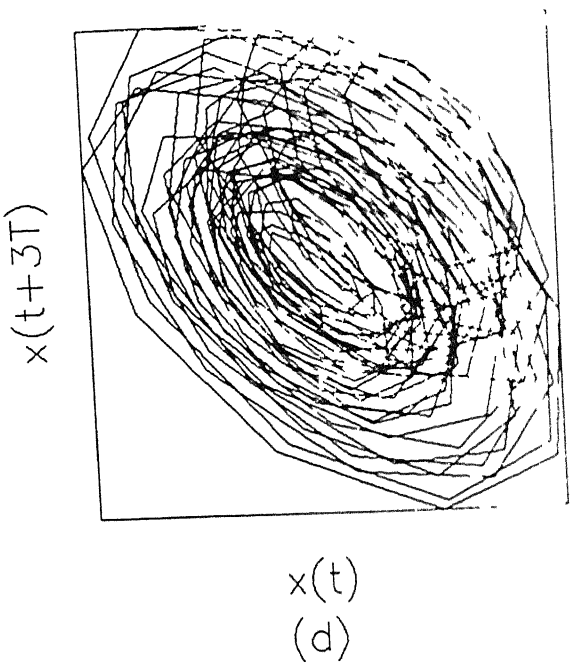
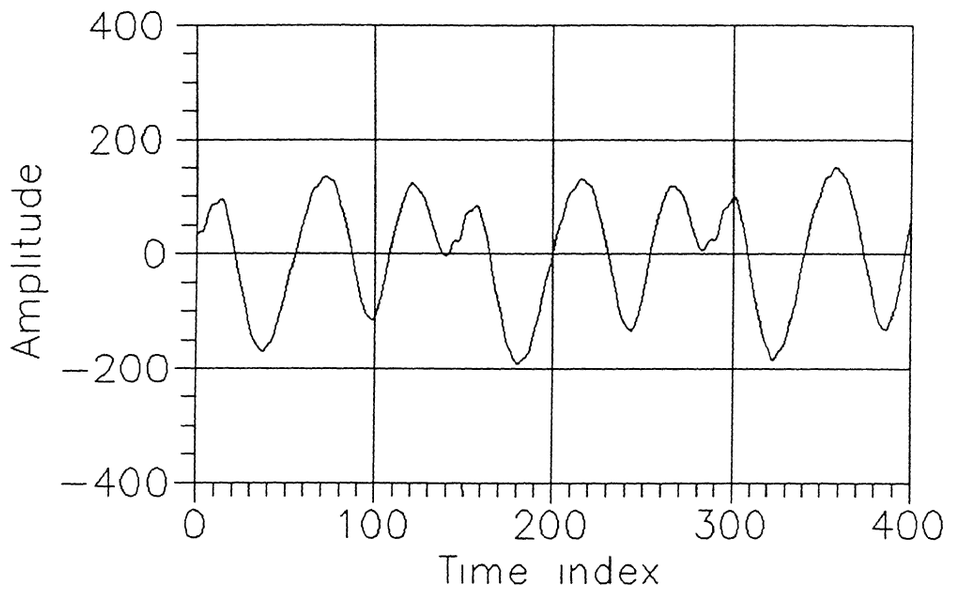
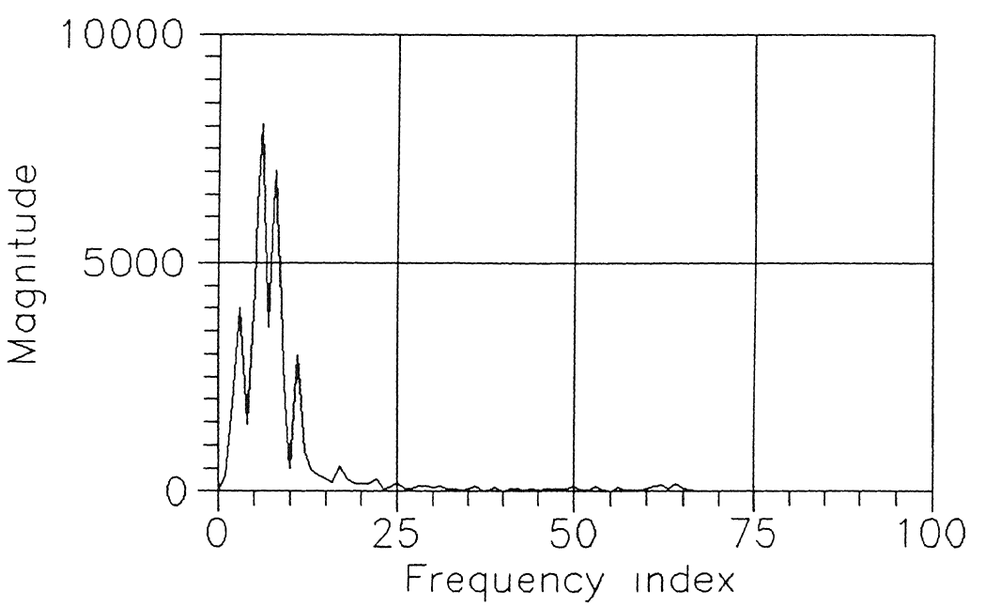


Fig. 2.6: For fricative utterance /ʒ/, (a) part of the time series at sampling rate $1/T$ = 16kHz, (b) the first 100 points of the corresponding 400 point fourier spectrum, (c) the projection of the reconstructed trajectory using SVD criterion on the 1-2 plane corresponding to the 2 largest singular values, (d) trajectory plot on the $(x(t), x(t+T_d))$ plane where T_d is estimated using minimum mutual information analysis


400
200
0
−200
−400
Amplitude
0
100
200
300
400
Time index

(a)


10000
5000
0
Magnitude
0
25
50
75
100
Frequency index

(b)

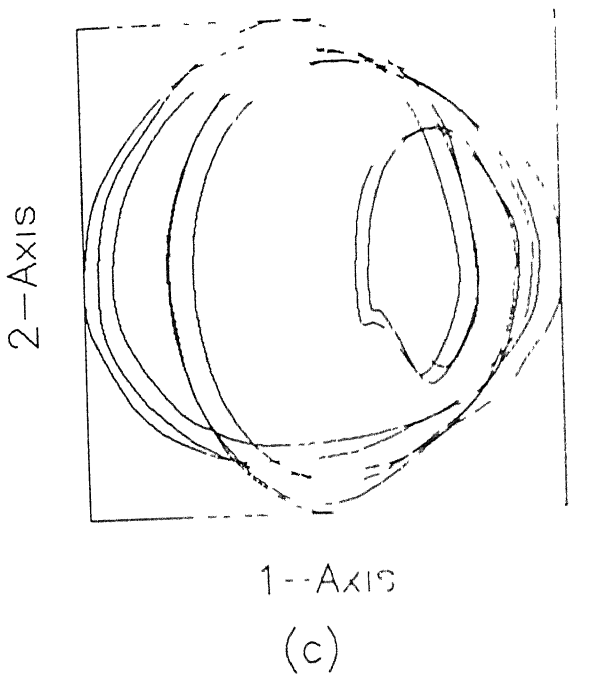


Fig. 2.7: For approximant utterance /l/, (a) part of the time series at sampling rate $1/T$ = 16kHz, (b) the first 100 points of the corresponding 400 point fourier spectrum, (c) the projection of the reconstructed trajectory using SVD criterion on the 1-2 plane corresponding to the 2 largest singular values, (d) trajectory plot on the $(x(t), x(t+T_d))$ plane where T_d is estimated using minimum mutual information analysis

2.3 Optimal State Space Reconstruction using Redundancy Criterion

This method is based on the requirement of general independence of the scalar variables of the reconstructed state space as compared to that of uncorrelatedness of the SVD method As an illustration, first consider the case of 2-dimensional reconstructed vectors,

$$\mathbf{x}^2(t+iT) = [x(t+iT)\ x(t+iT+T_d)]^T$$
$$\text{or} \quad \mathbf{x}_i^2 = [x_i\ x_{i+k}]^T \tag{2 9}$$

i e , $n = 2$ in eq (2 4) The parameter of interest is the time delay $T_d = kT$, in seconds, between the two variables If T_d is chosen to be very small, the two variables would convey nearly the same information and the reconstructed trajectory would lie close to the diagonal $x_i = x_{i+k}$ It is suggested that T_d is chosen as the first local minimum of the mutual information between them [46]

The mutual information between two random variables Y and Z is a functional defined by their joint probability density

$$I(Y,Z) = \left\langle \log_2 \left(\frac{p_{Y,Z}}{p_Y\ p_Z} \right) \right\rangle$$
$$= \int_{Y,Z} p_{Y,Z}(y,z) \log_2 \left(\frac{p_{Y,Z}(y,z)}{p_Y(y)\ p_Z(z)} \right) dy\ dz \tag{2 10}$$

where p_Y and p_Z are the probability density functions of Y and Z respectively, p_{YZ} is their joint probability density function and $<\ >$ denotes the ensemble average The logarithm is generally taken in base 2 so that the mutual information can be expressed in bits It can be shown that $I(Y,Z) = I(Z,Y)$ Also, $I(Y,Z) \geq 0$ If Y and Z are independent, i e , $p_{Y,Z} = p_Y\ p_Z$ then $I(Y,Z) = 0$ If on the other hand, z determines y exactly, then $I(Y,Z) = \infty$. Thus, mutual information between two variables is a measure of the general dependence between them In the above example, the variables are time delayed versions of the same scalar observable It answers the question, "Given $x(t)$, how many bits on the average can be predicted about $x(t+T_d)$?" Since we have a deterministic hypothesis on the data, strictly speaking, the mutual information between the two observables should be infinite

The divergence is avoided due to the finite accuracy of measurements and the use of finite partitions of the state space which are reflected in the result

The above concept of mutual information can be generalized to n random variables The *redundancy* between n random variables $Y_1, Y_2, \quad , Y_n$ can be defined as a functional of their joint probability density [45], [44]

$$R(Y_1, Y_2, \quad , Y_n) = \left\langle \log_2 \left(\frac{p_{Y_1, Y_2, \quad , Y_n}}{p_{Y_1}\, p_{Y_2} \quad p_{Y_n}} \right) \right\rangle$$
$$= \int_{Y_1, \quad Y_n} p_{Y_1, \quad , Y_n}(y_1, \quad , y_n) \log_2 \left(\frac{p_{Y_1, \quad , Y_n}(y_1, \quad , y_n)}{p_{Y_1}(y_1) \quad p_{Y_n}(y_n)} \right) dy_1 \quad dy_n \tag{2 11}$$

with the standard interpretation of the symbols It is possible to extend the minimum mutual information type of analysis to determine good reconstructions in n-dimensions using the concept of redundancy From eq (2 4), treating the n scalar variables as random variables, we have

$$Y_i = x(t + (i-1)kT)$$
$$= x(t + (i-1)T_d), \quad i = 1, \quad , n, \tag{2 12}$$

where the time delay $T_d = kT$ In this case, let us denote the redundancy in compact notation by $R^n_{T_d}$, where n is the embedding dimension and T_d, the time delay Redundancy gives a quantitative measure of the degree of dependence between the n time delayed scalar variables It is also helpful to define a quantity called *marginal redundancy*,

$$R'^n_{T_d} = I(x(t + nT_d), \mathbf{x}^n(t))$$
$$= R^{n+1}_{T_d} - R^n_{T_d}$$
$$= \left\langle \log_2 \left(\frac{p_{x(t+nT_d)/\mathbf{x}^n(t)}}{p_{x(t+nT_d)}} \right) \right\rangle \tag{2 13}$$

This tells us how many bits of the last component of $\mathbf{x}^{n+1}(t)$ can be predicted, on the average, from the previous n components

The idea of a good reconstruction is that a point on the reconstructed state space, $\mathbf{x}^n(t)$, provide as much *useful* information about the original state space point $s(t)$, as possible (see eq (2 3),(2 4)) That is, the conditional probability density $p_{s(t)/\mathbf{x}^n(t)}$ is

as sharp as possible Since the state space corresponding to the original dynamics is generally not accessible in experiments, the above criterion can heuristically be changed to the requirements that the conditional density $p_{x(t+nT_d)/\mathbf{x}^n(t)}$ be "as sharp as possible" and the scalar variables of $\mathbf{x}^n(t)$ be "as independent as possible" [107] A suboptimal algorithm to implement the above criteria may be arrived at in two successive steps

1. Choose an embedding dimension n so as to maximize the marginal redundancy $R'^n_{T_d}$. This ensures that $p_{x(t+nT_d)/\mathbf{x}^n(t)}$ is "sharply peaked".
2. Choose the appropriate time delay T_d as the first local minimum of the plot of redundancy $R^n_{T_d}$ versus time delay T_d

This algorithm is somewhat at variance with that originally proposed in [45] In that paper, it is argued that, with increasing scalar measurements, new information about the original system dynamics may get shifted from macroscales to microscales and, hence, be irrelevant Based on this argument it is stated that redundancy by itself does not represent the *useful* information about the dynamics For example, if the dynamics is chaotic,

$$\lim_{T_d\to\infty} p_{\mathbf{x}^n(t)} = \prod_{i=1}^{n} p_{x((n-i)T_d)} \tag{2 14a}$$

$$\lim_{T_d\to\infty} R^n_{T_d} = 0 \tag{2 14b}$$

Thus, minimizing the redundancy would mean choosing the largest possible value of T_d However, this is not an appropriate choice because it relates disparate length scales in the original system and the reconstructed dynamics As a remedy, a statistic is suggested in [45] that discounts for useless information regarding small scale features However, as a good first approximation, we endorse the choice of T_d as the first local minimum of the redundancy plot, as argued in [95]

We have used the minimum mutual information criterion to find the optimal time delay T_d for plotting 2-d reconstructed trajectories in state space A similar study using the redundancy criterion on sustained vowel utterances [a ,e ,i ,o ,u] by 3 male speakers is reported in [14] Our analysis was done on the same database as used in section 2 2 The mean time delay across 4 speakers varies from 0 19 ms for

fricatives /X/ and /h/ to 1 01 ms for nasal /ξ/ The time delay T_d over 44 consonants of the IPA (excluding the 13 plosives) across 4 speakers is obtained as $0\ 56 \pm 0\ 27$ ms Figures 2 3–2 7, part (d) show the 2-d reconstructed trajectories in the $x(t) - x(t + T_d)$ plane for three cardinal vowels /i/, /o/ and /a/, one fricative /ʒ/ and one approximant /ɻ/ respectively The following observations of this analysis and reconstruction scheme are noteworthy:

1 The mean time delay for the 8 Daniel Jones cardinal vowels spoken 4 times each is 0 55 ms while that for 22 fricatives across 4 speakers is 0 42 ms The difference in T_d between cardinal vowels and fricatives is intuitively expected because the *regular* nature of the former allows it a greater *predictability time* compared to fricatives

2 The variation in T_d across four repeated utterances of the same cardinal vowel and across four speakers of the same phoneme is relatively small The variation in T_d is less than 0 13 ms across 4 utterances of the same cardinal vowel for 7 of the 8 vowels. Similarly, the variation in T_d is less than 38 ms across utterances of the same phoneme by 4 speakers (3 males and 1 female) in the case of 33 of the 44 phonemes of the IPA that were analysed.

3 A comparative study of the reconstructed trajectories of the relatively *regular* cardinal vowels (fig 2 3–2 5, parts (c) and (d)), shows that there is little to distinguish between the two reconstruction schemes in this case However, in the case of more complex fricatives, the redundancy criterion appears to give consistently better reconstructions as far as visual comprehension is concerned, compared to the SVD criterion This fact is better appreciated if one observes the evolution of the reconstructed trajectories, say, on a computer screen We see in fig. 2 6(d), for fricative utterance /ʒ/, the reconstructed trajectory reproduces the periodicity corresponding to the prominent frequencies in the fourier spectrum of the utterance This periodicity will be visible in the SVD method also (fig 2 6(c)) if we view along the 1-axis rather than orthogonal to the 1–2 plane

The SVD method chooses a *good* reconstruction from a larger set of possible reconstructions compared to the mutual information method This is because the SVD

method gives an optimum basis set for reconstruction in addition to an appropriate time delay It also has an inbuilt ability for noise reduction. However, the mutual information method, which uses the criterion of statistical independence gives better reconstructions for specific examples [44] This is, however, not so in general [26]

After reconstructing the trajectory from a scalar time series, we will now discuss the method of Lyapunov exponents which is used to characterize the steady state behaviour of dynamical systems We will also give results of the estimation of the *largest* Lyapunov exponent from reconstructed speech trajectories

2.4 Lyapunov Exponents

Lyapunov exponents (also called *characteristic exponents*) give a coordinate independent measure of the local stability properties of a trajectory If the trajectory evolves in an n-dimensional state space, then the characterization is done through n exponents, usually arranged in decreasing order, and referred to as the "spectrum of Lyapunov exponents" Stability characterizes response to perturbations and can refer to individual points (local stability), to trajectories (asymptotic local stability), to families of trajectories (attractors) or to an entire dynamical system (whether there is a unique attractor or not). Conceptually, Lyapunov exponents are a generalization of eigenvalues used in characterizing different types of equilibrium points They categorize bounded trajectories into equilibrium points, periodic solutions, quasiperiodic solutions or chaotic solutions The first three types of steady state trajectories are called *regular* trajectories They are asymptotically locally stable and are characterized by a spectrum of nonpositive Lyapunov exponents An equilibrium solution has a Lyapunov exponent spectrum of the form $(-,-,\ )$ A limit cycle (example of a periodic solution) has a spectrum of the form $(0,-,-,\ )$, while a two-torus (example of a quasiperiodic solution) has a spectrum of the form $(0,0,-,-,\ )$ In contrast, a chaotic solution which is asymptotically locally unstable, is defined by the presence of at least one positive Lyapunov exponent.

2.4.1 Theory and Evaluation from Scalar Time Series

Let us first consider the case where the dynamical equations are explicitly known [36], [50] We take a discrete time dynamical system

$$s_{k+1} = f(s_k) \tag{2 15}$$

where f $R^n \rightarrow R^n$ is a differentiable vector function Consider the growth of an arbitrary small perturbation Δs_0 about a point s_0 The one step evolution of this perturbation can be approximated as follows,

$$\begin{aligned} s_1 &= f(s_0) \\ s_1 + \Delta s_1 &= f(s_0 + \Delta s_0) \\ \Delta s_1 &\simeq (\partial_{s_0} f)\ \Delta s_0 \\ &= J_s\ \Delta s_0 \end{aligned} \tag{2 16}$$

where J_s is the Jacobian of f Each of the eigenvectors of J, locally grow at a rate e^{λ_j}, where $\lambda_j, j = 1, 2, \quad , n$ are the eigenvalues of the matrix The growth of the perturbation after i steps is given by the i^{th} iterate f^i of f, and can be approximated by its Jacobian, J_s^i Using the chain rule,

$$\begin{aligned} \Delta s_i &\simeq J_s^i\ \Delta s_0 \\ &= (\partial_{s_0} f^i) \Delta s_0 \\ &= (\partial_{s_{i-1}} f)(\partial_{s_{i-2}} f) \quad (\partial_{s_0} f) \Delta s_0 \end{aligned} \tag{2 17}$$

The Lyapunov exponents are given by the asymptotic growth rate of the eigenvalues of J_s^i

$$\lambda_j = \lim_{i \rightarrow \infty} \frac{1}{i} \log \| J_s^i \|, \ j = 1, \quad , n \tag{2 18}$$

The existence of the limit is assured by the multiplicative ergodic theorem of Oseledec [106] The above definition can also be extended to continuous time dynamical systems

Lyapunov exponents can thus be calculated by linearizing the system of equations and applying it to small perturbations about a given numerical solution When the system equations are not known, it is possible to get the Lyapunov exponents from a scalar observable of the evolution of the dynamical system This consists of reconstructing the state space, following the evolution of the reconstructed trajectories that are close to each other and fitting the data in local neighbourhoods to approximate J_s^i The reader is referred to [35], [119] for algorithmic details While in those papers, linear maps are used, a later work [22] shows that it is advantageous to use higher

order Taylor series for local mappings We have, however, used a different algorithm for computing the largest Lyapunov exponent from speech time series. This is because the algorithm is more robust to changes in input parameters In contrast, the previous algorithm has been found to yield Lyapunov exponent estimates that vary considerably with the embedding dimension [147] To discuss this algorithm, let us consider a more formal definition of Lyapunov exponents [63]

Let f $R^n \rightarrow R^n$ define a discrete dynamical system Assume f is C^1 Fix s $\in R^n$ Let $f^i(s)$ denote the ith iterate of f Suppose that there are subspaces $V_i^{(1)} \supset V_i^{(2)} \supset \quad \supset V_i^{(n)}$ in the tangent space of $f^i(s)$ and numbers $\lambda_1 \geq \lambda_2 \geq \quad \geq \lambda_n$ that depend on s with properties that

(a) $\partial f(V_i^{(j)}) = V_{i+1}^{(j)}$, $i = 0, 1, \quad , j = 1, 2, \quad , n$ (2 19a)

(b) $dim V_i^{(j)} = n + 1 - j$, $i = 0, 1, \quad , j = 1, 2, \quad , n$ (2 19b)

(c) $\lim_{n\rightarrow\infty} \frac{1}{n} \log \|\partial_s f^n(s^j)\| = \lambda_j$, for all $s^j \in V_0^{(j)} - V_0^{(j+1)}$, $\|s^j\| = 1$ (2 19c)

then $\lambda_j, j = 1, \quad , n$ are the n Lyapunov exponents of f

As an example, consider for $n = 2$,

$$\partial_s f = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix} \tag{2 20}$$

where, s is a fixed point and $a_1 > a_2$. Pick $V^{(1)} \equiv span[(0,1),(1,0)\} = R^2$, and $V^{(2)} \equiv span[(0,1)\}$ Then $V_i^{(1)} = V^{(1)}, V_i^{(2)} = V^{(2)}, i = 0, 1,$ satisfy (a) and (b) Also, $\lambda_j = \log a_j, j = 1, 2$ In general, $V_0^{(1)} - V_0^{(2)}$ consists of all vectors which grow at the fastest possible rate, $V_0^{(2)} - V_0^{(3)}$ consists of all vectors which grow at the next fastest rate and so on. Alternatively, $V_0^{(1)}$ is the direct sum of eigenspaces corresponding to $\lambda_i, i = 1, \quad , n$ $V_0^{(2)}$ is the direct sum of eigenspaces corresponding to $\lambda_i, i = 2, 3, \quad , n$, where $\lambda_1 > \lambda_2$ and so on [10] Since $dim V_0^{(1)} = n$, if one chooses the vector s^j in (c) "randomly", it is most likely to lie in $V_0^{(1)} - V_0^{(2)}$ which is of full Lebesgue measure. This vector will evolve at a rate given by the largest Lyapunov exponent. Similarly, an area element defined by three points in the state space would evolve at a rate proportional to $e^{(\lambda_1+\lambda_2)n}$ A d-dimensional volume element would grow at a rate depending on the first d eigenvalues This gives a method for estimating the Lyapunov exponents If the system equations are known, then

one can monitor the evolution of a set of small, initially orthogonal vectors (called *principal axis vectors*) about a central trajectory (called the *fiducial trajectory*) The growth rate of a single principal axis vector gives λ_1, the growth rate of the area spanned by two vectors gives $\lambda_1 + \lambda_2$ etc. The exponential dominance of the largest eigenvalue requires that the principal axes be periodically reorthogonalized so that they remain resolvable in the finite precision of computers

The above discussion suggests a procedure to evaluate the largest Lyapunov exponent from a scalar observable of a dynamical system For algorithmic details, its implementation aspects and justifications, see [149] Here, we confine to enumerating the steps of the algorithm Given time series $x_i, i = 1, \ldots, N$, following are the steps (see also fig 2 8)

STEP 1: Reconstruct the trajectory in n-dimensional state space using time delay embeddings

$$\mathbf{x}_i^n = [x_i \; x_{i+k} \; \ldots \; x_{i+(n-1)k}]^T, \; i = 1, \ldots, N - (n-1)k \tag{2.4}$$

STEP 2: Start the algorithm by locating the nearest neighbour $\mathbf{x}_{t_1}^n$, whose distance is greater than a prefixed distance parameter SCALMN, to the initial point on the fiducial trajectory $\mathbf{x}_1^n$ Let $d_1^{(1)} = \|\mathbf{x}_1^n - \mathbf{x}_{t_1}^n\|$ SCALMN ensures that distances below the noise scale are not used in the computation

STEP 3: Set $d_2^{(1)} = \|\mathbf{x}_{1+q}^n - \mathbf{x}_{t_1+q}^n\|$, where q is the prefixed number of steps for which the pair of points are to be evolved Store $l_1 = d_2^{(1)}/d_1^{(1)}$

This completes the initial run Now enter the main loop

STEP 4: A replacement point $\mathbf{x}_{t_2}^n$ is found as follows The distance between the replacement point $\mathbf{x}_{t_2}^n$ and the fiducial point $\mathbf{x}_{1+q}^n$, i e , $d_1^{(2)} \quad \|\mathbf{x}_{1+q}^n - \mathbf{x}_{t_2+q}^n\|$ is the smallest distance greater than SCALMN, and less than a prefixed parameter SCALMX to ensure that the initial principal axis vector is not large Also, the angular orientation θ_1 between the replacement vector and the evolved vector is less than a prefixed value ANGLMX

STEP 5: If no points satisfy the criteria in STEP 4, relax the parameter SCALMX and repeat STEP 4 until such a point is found

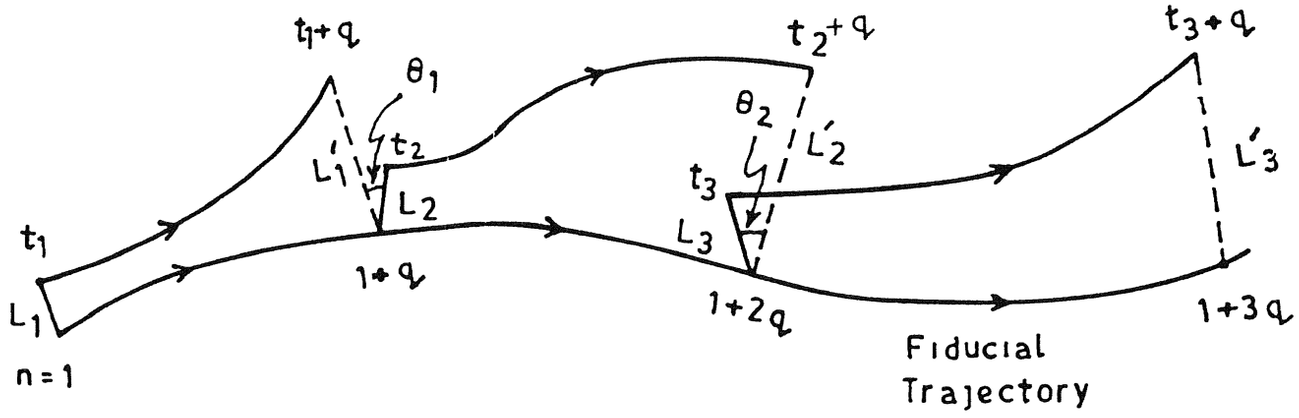


Fig. 2.8: Schematic showing the estimation of the largest Lyapunov exponent from time series In the reconstructed state space, the evolution of a principal axis vector along the fiducial trajectory is monitored After an evolution of q steps, a new principal axis vector of minimum length L_i is chosen such that the orientation angle θ_i is less than a prefixed value

STEP 6: Set $d_2^{(2)} = \|\mathbf{x}^n_{1+2q} - \mathbf{x}^n_{t_2+q}\|$ Store $l_2 = d_2^{(2)}/d_1^{(1)}$

Repeat steps 4–6. In general, for the i^{th} iteration,

$$d_1^{(i)} = \|\mathbf{x}^n_{1+(i-1)q} - \mathbf{x}^n_{t_i}\| \tag{2.21a}$$

$$d_2^{(i)} = \|\mathbf{x}^n_{1+iq} - \mathbf{x}^n_{t_i+q}\| \tag{2.21b}$$

$$l_i = \frac{d_2^{(i)}}{d_1^{(i)}} \tag{2.21c}$$

Continue till $i = N - (n-1)k = N'$ (say) The estimate of the largest Lyapunov exponent is then given by

$$\hat{\lambda}_1 = \frac{1}{N'q} \sum_{i=1}^{N'} \log\left(\frac{d_2^{(i)}}{d_1^{(i)}}\right) \tag{2.22}$$

That this definition indeed gives the largest Lyapunov exponent can be seen from the fact that the choice of replacement vectors come from the full measure set $V_0^{(1)} - V_0^{(2)}$ It is necessary to periodically replace the principal axis vector so as to prevent it from going through a global "fold" when we are only interested in averaging the local "stretch" The parameters that need to be fixed before running the algorithm are the embedding dimension n, time delay k, SCALMN, SCALMX, ANGLMX and the number of evolution steps q Although fixing the parameter values requires some experience with the algorithm, the estimate is usually robust to a reasonably large range of parameter variation

Estimating negative Lyapunov exponents from time series is difficult because it is often impossible to resolve information about the contracting state space directions Volume elements involving negative exponent directions collapse exponentially fast For most applications, finding the largest Lyapunov exponent is sufficient because a positive value indicates the sensitive dependence of trajectories on the initial condition

2.4.2 Results from Speech Signal and Some Comparisons

We have used the above algorithm to estimate the largest Lyapunov exponent from phonemes The speech database and the ADC parameters are as given in database 1 (Appendix B) Again, we excluded plosives from this analysis because of the highly

nonstationary nature of any meaningful length of the corresponding time series The exponent was computed for all the other phonemes using the algorithm described in section 2 4 1 The algorithm was run several times on each time series using different sets of parameter values The results obtained are fairly stable over a large range of parameter values The data length used for computation is $N = 2000$ In all instances, the Lyapunov exponent converged in far fewer iterations than afforded by the data length Figure 2 9 shows the convergence of the largest Lyapunov exponent as a function of the number of iterations for two phonemes — /δ/, $\lambda_1 = 4161\ 7s^{-1}$ and /Ş/, $\lambda_1 = 2038\ 1s^{-1}$ The time series in each case consists of 2000 samples An evolution step, $q = 5$ per iteration allows approximately 400 iterations

The largest Lyapunov exponent gives an indication of the steady state behaviour of the dynamical system The results of the largest Lyapunov exponent estimation are summarized in Table 2.1 The error values indicate the standard deviation (s d) of the mean value of the Lyapunov exponent per phoneme type over four speakers In the case of cardinal vowels, the s d is over four utterances by a single speaker The positive values give evidence of the exponential divergence of nearby trajectories

Phoneme Type	No. of Phonemes	Largest Lyapunov Exponent (s^{-1})
Cardinal Vowel	8	3482 2 ± 182 5
Nasal	7	2198 9 ± 700 3
Trill	2	3742 7 ± 1382 4
Tap or Flap	2	4469 9 ± 1009 4
Fricative	22	2597 3 ± 854 9
Lateral Fricative	2	2105 9 ± 774 3
Approximant	5	3345 4 ± 1191 1
Lateral Approximant	4	3247 2 ± 725 7
		Mean = 2899 0

Table 2 1: Mean value of the largest Lyapunov exponent for various phoneme types The error values indicate the s d of the mean per phoneme type over 4 speakers In the case of cardinal vowels, the s d is over 4 utterances by a single speaker

To place more faith in the algorithm and results, we generated three time series with known dynamics and used the same algorithm to evaluate the largest Lyapunov exponent The three equations are

$$y_1(t) = 10\sin 250\pi t \tag{2.23a}$$

$$y_2(t) = 3\sin 250\pi t + 7\sin \frac{500}{3}\pi t + 4\sin 125\pi t \tag{2.23b}$$

$$y_3(t) = (\sin 50\pi t)\ \ (\sin 50\sqrt{11}\pi t) \tag{2.23c}$$

$y_1(t)$ and $y_2(t)$ are periodic with one and three frequencies respectively while $y_3(t)$ is quasiperiodic with two incommensurate frequencies The three waveforms are sampled at 16 kHz to produce three time series $y_1(i), y_2(i)\ y_3(i), i = 1, \ldots, 2000$. Ideally, all the three time series should produce zero Lyapunov exponents Using the above algorithm, the computed exponents are $152.8 \pm 83.3, 50.9 \pm 19.6$ and 33.1 ± 38.2 respectively The error values show the variation with respect to the algorithm parameter values

To compare the results with those of speech signals, we now mimic the conditions on $y_i(t), i = 1, 2, 3$, under which the speech signals were analysed The three time series are multiplied by appropriate gain factors and quantized to 12 bits/samples The exponent for the three quantized time series are obtained as $74.7 \pm 250.2, 61.0 \pm 41.5$ and 294.0 ± 89.3 respectively Next consider the effect of observational noise Noise with uniform density is generated using a one step whitening filter and added to the time series prior to quantization Five cases of noise s d equivalent in magnitude to the lower 0 to 4 bits respectively are considered The corresponding largest Lyapunov exponent of the three time series are shown in fig 2.10 It is seen that the exponent value increases with noise variance

In all instances of speech signal analysis, the Lyapunov exponent is greater than $600s^{-1}$ For synthesised data, barring the unrealistic instances of noise s d equal to 3 and 4 bits, the exponent is significantly less than $600s^{-1}$, and is occasionally negative Moreover, the mean for speech signals is one to two orders of magnitude greater than that for computer generated periodic and quasiperiodic time series, even in the presence of significant additive noise in the latter cases Hence, Lyapunov exponent analysis shows that reconstructed speech trajectories can be distinguished from

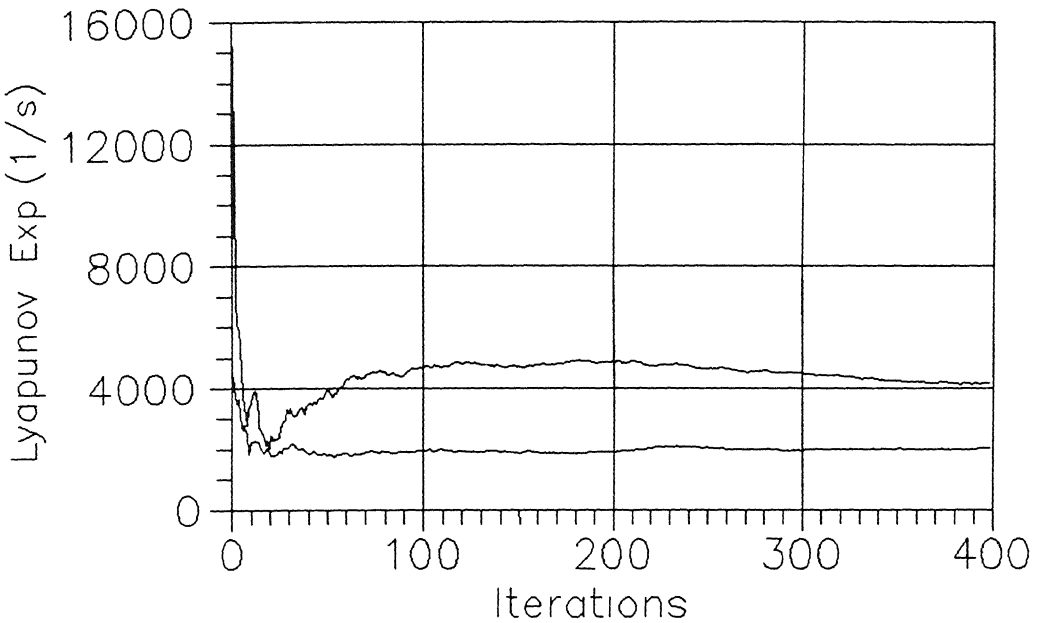


Fig. 2.9: The convergence plot of the largest Lyapunov exponent as a function of the number of iterations for two phonemes

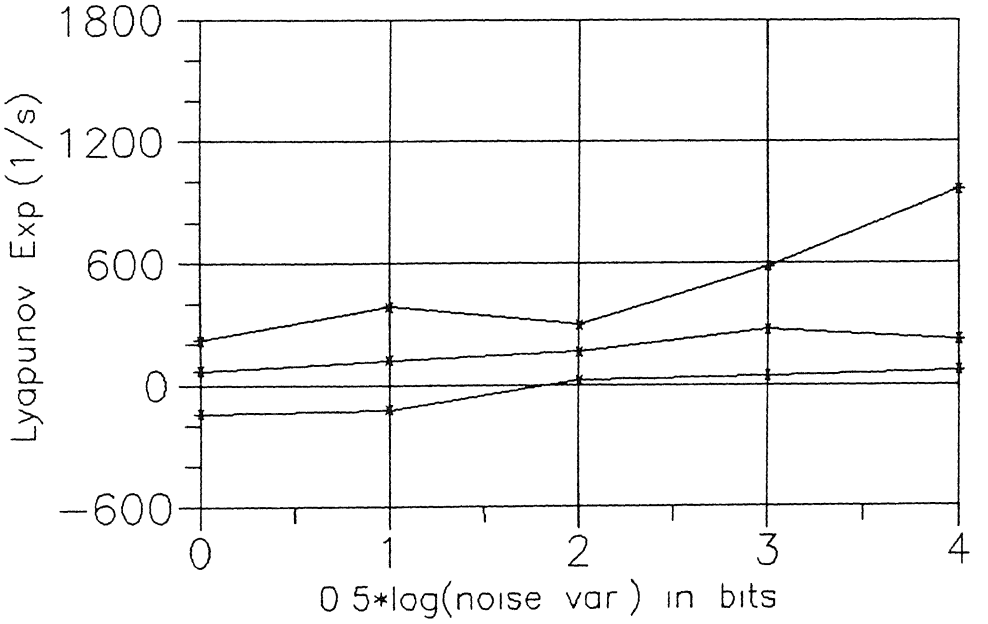


Fig. 2.10: The largest Lyapunov exponent for 3 time series generated from eq 2.23 (a)–(c) as a function of the variance of additive noise The bottom to top plots correspond to eq 2.23(a), (b) and (c) respectively

periodic and quasiperiodic behaviour in that nearby speech trajectories exhibit exponential divergence on the average The relation between Lyapunov exponents and other dynamical invariants, namely dimension and metric entropy, will be discussed in the next chapter

Chapter 3

Dynamical Analysis of Speech Signals—2

Dimension and metric entropy are two important parameters characterizing the behaviour of dynamical systems The state space reconstruction theorems of Takens, discussed in chapter 2, allow us to get an estimate of these "invariants" from a scalar observable of the time evolution of a dynamical system This becomes important when one wants to build *nonlinear* state space models of observed phenomena In this chapter, we review the definitions of dimension and dynamical entropy and their estimation procedures from time series data We also give results of our estimation of these invariants for unit speech utterances, namely phonemes, and draw conclusions from them Towards this end, the chapter is organized as follows In section 3 1, we dwell qualitatively on the notions of dimension and entropy A discussion of the concepts of invariant and natural measures of a dynamical system in section 3.2 builds a ground for the definitions of dimension and dynamical entropy in sections 3.3 and 3 4 respectively We also consider the estimation of these dynamical invariants from the generalized correlation sum in these respective sections Based on the developments in sections 3 3 and 3 4, we review a method for the joint estimation of the correlation dimension and second order entropy from time series in section 3 5 Section 3 6 deals with the estimation of correlation dimension from speech time series We give results of the numerical computation from speech time series as also

from a simplified model of a specific vowel utterance In section 3 7, we discuss some implementation aspects of the correlation sum algorithm used in the estimation of the invariants We also consider how various sources of error in estimation affect the dimension results Finally, in section 3 8 we present the results of the computation of second order entropy from speech time series and draw conclusions from them

3.1 Notion of Dimension and Entropy

Various definitions of dimension and entropy are used to characterize, both theoretically and experimentally, the steady state behaviour of dynamical systems The notions of dimension and entropy have been given a wider interpretation in the last few decades to quantify dynamical systems behaviour, and specially the complex behaviour of chaotic dynamics

Dimension is one of the basic properties associated with the geometrical description of an object An unqualified reference to the dimension of an object is usually made to the Euclidean space in which it resides Dimension is thus associated with the "minimum number of independent directions" of the space containing the object One can also look upon the dimension of an object as the exponent according to which its "volume" or "bulk" scales with the resolution For example, the volume of a cube scales as the third power of its side In dynamical systems literature, dimension is used to specify the number of degrees of freedom that a system possesses [137], [37], [61], [50] It is useful here to distinguish between *nominal* and *effective* degrees of freedom Nominal degrees of freedom refers to the number of state variables needed to describe the dynamical system Although there may be many nominal degrees of freedom, the dynamics may settle down on or approach a subset of the state space, called the *attractor* Attractors exist only for dissipative dynamical systems and can be any one of four types, namely, fixed point, limit cycle, q-torus in the case of quasiperiodicity and strange attractor in the case of chaotic dynamics (see Appendix B and references therein) The dimension of these objects on which the steady state dynamics settles down gives the number of effective degrees of freedom The dimension of the first three types of attractors are integral numbers Chaotic dynamics produces complex trajectories which asymptotically settle down on strange

attractors These attractors are usually *fractal* or self-similar and the scaling exponents of their appropriate measures with respect to the resolution are fractional or non-integral Thus, strange attractors generally have nonintegral dimension and various definitions exist to precisely give the number of effective degrees of freedom Given a time series as a scalar projection of a state space trajectory, dimension analysis can be used to quantify its complexity It gives the *necessary* and *sufficient* number of state variables needed to approximate the steady state dynamics from which the time series is realized. Since dimension calculation from time series gives the number of effective degrees of freedom in the underlying dynamics which is usually much smaller than the number of nominal degrees of freedom, a phenomenological state space modelling scheme of the time series can inherently eliminate any redundancy of variables

Kolmogorov or metric entropy is another important measure of dynamical systems by which their time evolution can be characterized [125], [50] Historically, entropy has thermodynamic origins where it is used as a measure of the disorder in a given system The increase in disorder of a system is associated with an increase in the uncertainty of the state of the system at a given instant of time In information theoretic terms, entropy of a system denotes the average uncertainty or the amount of information required to locate the system in a specified state Metric entropy quantifies the "degree of chaos" of a dynamical system When a dynamical system is under observation, each new measurement is a potential source of additional information about the system For example, in the case of fixed point behaviour, no additional information is obtained after a single measurement in the steady state Similarly, in the case of periodic and quasiperiodic behaviour, no additional information is obtained after a finite number of measurements In contrast, a chaotic dynamical system continues to produce new information with each succeeding measurement The metric entropy, K, is an asymptotic measure of the average rate of information production as a function of time Thus, $K = 0$ for situations of *regular* behaviour (fixed point, limit cycle, quasiperiodicity), $0 < K < \infty$ for chaotic behaviour and $K = \infty$ for truly random behaviour When the laws governing the dynamical system or a phenomenological state space model of the observed time series is available,

metric entropy can also be looked upon as the asymptotic rate at which information about the initial condition driving the system is lost In other words, metric entropy is proportional to the inverse *predictability time* of the evolution of a dynamical system from a given initial condition

The theory of state space reconstruction discussed in chapter 2 allows us to reconstruct trajectories in the state space from a scalar time series (or *observable*) of the dynamical system evolution such that the dynamical invariants evaluated from the reconstructed trajectory are the same as those of the underlying dynamical system This means that it is possible to estimate dimension and metric entropy (also Lyapunov exponents, Chapter 2) of a stationary vocal tract configuration from the corresponding speech signal In the next section, we discuss the concepts of invariant and natural measures of dynamical systems These are relevant in the context of the definitions of dimension and dynamical entropy

3.2 Invariant and Natural Measures of a Dynamical System

The characterization of dynamical systems behaviour through dimensions and entropy can be done either from a physical or phenomenological model of the dynamics or from an observation of its time evolution. This is because dimensions and entropy are estimated from the *invariant* probability measure of the dynamical system which can be analytically obtained from the physical model of the dynamics or estimated from its time evolution by invoking ergodicity [36], [37], [137], [132] To discuss this further, let us consider a discrete time dynamical system $\mathbf{f}\ \ M \rightarrow M$, where M (usually R^n) is the state space on which the map evolves from an initial condition $\mathbf{s}_0$

$$\mathbf{s}_{i+1} = \mathbf{f}(\mathbf{s}_i), \quad i = 0, 1, \tag{3 1}$$

In the case of continuous time dynamical systems, assume that a corresponding discrete time map $\mathbf{f}$ exists by discretising time or Poincare sectioning [125] In experimental situations, one can only observe the time evolution of one or at most a few variables of the dynamical system $\mathbf{f}$ Consider a situation where just a scalar time

series x_i, $i = 0, 1,$ is observed through the smooth *observable* function $h \quad M \rightarrow R$ on the dynamical system That is,

$$\begin{aligned} x_i &= h(\mathbf{s}_i) \\ &= h\left(\mathbf{f}^i(\mathbf{s}_0)\right), \quad i = 0, 1, 2, \end{aligned} \tag{3 2}$$

One can talk of dimensions, metric entropy and Lyapunov exponents of a single trajectory of a dynamical system However, it is more meaningful to associate these terms with a statistical description of the dynamics In regular and chaotic dynamical systems, there exist sets of initial conditions which give rise to families of trajectories having the same statistical properties, for example, when different trajectories of a dissipative dynamical system asymptotically settle on the same attractor A geometrical description of attractors, particularly strange attractors, presents mathematical difficulties Shifting the attention to a description through invariant measures simplifies the proceedings

An *invariant measure* is one that does not change under the action of the dynamics That is, a measure μ is invariant under the map $\mathbf{f}$, if for any subset S of M which is in the support of μ,

$$\mu(S) = \mu(\mathbf{f}^{-i}(S)) \tag{3 3}$$

for all $i > 0$ Here, $\mathbf{f}^{-i}(S) \equiv \{\mathbf{s} \in \mathbf{M} \quad \mathbf{f}^i(\mathbf{s}) \in \mathbf{S}\}$ In other words, a measure of a set is invariant if it is equal to that of the sets mapped into it Lebesgue measure, which is one of the most common measures, is not invariant with respect to dissipative dynamical systems because state space volumes contract under the action of the dynamics Therefore, one has to look for other physically relevant invariant measures

Although a dynamical system may have many invariant measures, not all of them are relevant because they may not be physically observable Of particular importance is the (computer generated) time evolution or experimental observation of dynamical systems Operationally, they appear to produce well defined time averages This gives us a selection procedure for a measure μ through ergodicity. Such a measure

is called the *natural measure* of the dynamical system which is invariant with respect to the dynamics Thus, we have

$$\mu(S) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} I_S\left[\mathbf{f}^i(\mathbf{s}_0)\right] \tag{3 4}$$

where $I_S(\mathbf{s}) = 1$ if $\mathbf{s} \in S$ and 0 otherwise Also,

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \mathbf{g}(\mathbf{s}_i) &= \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \mathbf{g}(\mathbf{f}^i(\mathbf{s}_0)) \\ &= \int_M \mathbf{g}(\mathbf{f}(S)) d\mu(S) \end{aligned} \tag{3 5}$$

for a continuous function g(s) over the map f of the underlying dynamical system A dynamical system may have many natural measures corresponding to different families of trajectories In a dissipative dynamical system, for example, each attractor will have its own natural measure For an attractor, the natural measure is distinguished from other measures because it is stable to low level dynamical noise That is, if the deterministic dynamical system is replaced by one with some additive dynamical noise, as long as the noise is small enough, the natural measure of the noisy system will remain similar to that of the deterministic system

The natural measure can be estimated numerically from a time evolution of the dynamical system using a histogram or kernel density estimation procedure

The vocal tract system is a complex dynamical system and an accurate physical model must consider the effects due to time variation of the vocal tract shape, losses due to heat conduction and viscous friction at the vocal tract walls, radiation of sound at the lips, nasal coupling, the excitation function due to air flow from the lungs etc. [114], [124] These considerations suggest that an accurate physical model would be a nonlinear dissipative dynamical system which can possibly show chaotic behaviour The computation of dimension and metric entropy (also Lyapunov exponents) is a tricky business because of the time varying nature of the vocal tract and the simultaneous requirement of large data length and stationarity in the time interval in which the computation is done When we refer to dimension, metric entropy and Lyapunov exponents as "dynamical invariants", we assume that there

exists a stationary (or *invariant*) probability measure of the underlying attractor The numerical evaluation of these invariants from a single time evolution (or realization) of a dynamical system assumes ergodicity which is rarely proved even in those cases where a physical model of the dynamics is available The time varying nature of the vocal tract dynamics and the consequent nonstationarity of the speech signal requires us to compute dimension and metric entropy from approximately stationary segments of phoneme utterances. It is important to note that these parameters characterize the corresponding speech signal and their reconstructed trajectories Linking them to any underlying attractors is only hypothetical and not proved

3.3 Definitions of Dimension and Relation with the Generalized Correlation Sum

In dynamical systems theory, dimension is used to characterize the objects on which trajectories asymptotically accumulate As mentioned before, the intuitive notion of dimension is as a scaling exponent of the "bulk" or "volume" of the object with respect to the resolution, i e ,

$$bulk \sim (resolution)^{dimension} \tag{3 6a}$$

Here, *bulk* can refer to the volume, mass or some measure of the object The definition of dimension is usually made in the form

$$dimension = \lim_{resolution \to 0} \frac{\log bulk}{\log resolution} \tag{3.6b}$$

In dissipative dynamical systems, the steady state motion takes place on or asymptotically approaches an attractor which has Lebesgue measure zero. In this case, dimension is used to characterize the attractor on which typical trajectories accumulate, i e those which originate from a set of initial conditions having positive Lebesgue measure The relevant measure of the bulk or volume of the attractor $\mathcal{A}$ or any subset $\mathcal{B}$ is the natural invariant measure $\mu(\mathcal{A})$ or $\mu(\mathcal{B})$ associated with it The normalization of this measure for any subset $\mathcal{B}$ with respect to the attractor $\mathcal{A}$ will be the corresponding probability that a random point on the attractor lies in $\mathcal{B}$, i e ,

$$P_{\mathcal{B}} = \frac{\mu(\mathcal{B})}{\mu(\mathcal{A})} \tag{3 7}$$

For simple attractors characterizing regular motion, any reasonable definition of dimension gives the same number Thus, a fixed point has zero dimension, a limit cycle has dimension one and a q-torus has dimension q More complicated geometrical objects like chaotic attractors often exhibit fractal properties [125], [8] There exist a variety of definitions of dimension (often producing fractional numbers) to describe these objects. Not all of them give the same number. In fact, the differences between these dimensions can be used to characterize the fractal structure of strange attractors These definitions of dimension can also be used to characterize regular attractors

The origin of these definitions of dimension can be traced back to Hausdorff who gave a completely rigorous definition of dimension in 1919 The Hausdorff dimension utilizes a purely geometrical description of the possibly fractal set and makes no reference to the measure μ defined on the attractor [137], [39], [152] It is also not amenable to numerical estimates To define the Hausdorff dimension of a set $\mathcal{A}$ lying in a n-dimensional Euclidean space, consider a finite covering $C(r,\mathcal{A}) = \{\mathcal{B}_1, \mathcal{B}_2, \quad, \mathcal{B}_k\}$ of $\mathcal{A}$ with sets $\mathcal{B}_l, l = 1, \quad, k$ whose diameters r_l are less than r Define the quantity $l_d(r)$ by

$$l_d(r) = \inf_{C(r,\mathcal{A})} \sum_l r_l^d \tag{3 8}$$

where the infimum (or minimum) extends over all coverings satisfying the constraint $r_l \leq r$. Now let

$$l_d = \lim_{r \to 0} l_d(r) \tag{3 9}$$

Then there exists a critical value of d above which $l_d = 0$ and below which $l_d = \infty$ This critical value, $d = d_H$, is the Hausdorff dimension

The difficulty in estimating the Hausdorff dimension numerically is that an infimum over all coverings has to be taken If one relaxes this requirement and chooses a covering that uses fixed size boxes or a partition of side resolution r, then one obtains an upper bound to the Hausdorff dimension This is variously referred to as the capacity, the fractal dimension or the box counting dimension [137], [39],

[152] Thus,

$$\begin{aligned} l_d(r) &= \sum_l r_l^d \\ &= \sum_l r^d \\ &= n(r)r^d \end{aligned} \tag{3 10}$$

where $n(r)$ is the number of nonempty boxes of resolution r The capacity D_c is the value of d at the transition between $l_d = 0$ to $l_d = \infty$. Taking $l_d \sim 1$, gives

$$n(r) \sim r^{-D_c} \tag{3 11a}$$

or, formally,

$$D_c = \lim_{r \to 0} \frac{\log \frac{1}{n(r)}}{\log r} \tag{3 11b}$$

The definition of capacity only uses the information about the number of nonempty boxes It does not take into account the number of points in each box In other words, the underlying probability measure is ignored and a uniform probability for each box is assumed

The *generalized dimensions* [66], [54], [112] were proposed to give a unified approach to the various definitions of dimension existing at that time and explain the differences in them Generalized dimensions take into account the probability measure of each box covering the attractor Let $\Pi = [\pi_1, \pi_2, \quad, \pi_{M(r)}]$ denote a finite partition of side resolution r of the attractor $\mathcal{A}$. Here, $M(r)$ denotes the number of partitions of Π. Also let $P_{\pi_l} = \frac{\mu(\pi_l)}{\mu(\mathcal{A})}$ be the normalized probability measure of the partition π_l, $l = 1, \quad, M(r)$ The generalized dimensions D_q are defined by

$$D_q = \lim_{r \to 0} \frac{1}{q-1} \frac{\log \left(\sum_{l=1}^{M(r)} P_{\pi_l}^q \right)}{\log r} \tag{3 12}$$

for any $q \neq 1$ The value for $q = 1$ is defined by taking the limit as $q \to 1$ This is known as the *information dimension* [137], [39], and is given by

$$\begin{aligned} D_1 &= \lim_{q \to 1} D_q \\ &= \lim_{r \to 0} \frac{\sum_{l=1}^{M(r)} P_{\pi_l} \log P_{\pi_l}}{\log r} \end{aligned} \tag{3 13}$$

Of particular importance is D_2 which is called the *correlation dimension* [57], [58] It is given by

$$D_2 = \lim_{r \to 0} \frac{\log \left(\sum_{l=1}^{M(r)} P_{\pi_l}^2 \right)}{\log r} \tag{3 14}$$

It is one of the most popular dimension computation algorithms and can be easily used on experimental time series data via the *correlation sum* as discussed below

Similarly, it can be shown that D_0 is the capacity or the fractal dimension For a uniform probability measure on the attractor, D_q is the same for all q For a nonuniform attractor or *multifractal*, the variation in D_q with q quantifies its nonuniformity For example,

$$D_\infty = \lim_{r \to 0} \frac{\log \left(\max_l P_{\pi_l} \right)}{logr} \tag{3 15a}$$

$$D_{-\infty} = \lim_{r \to 0} \frac{\log \left(\min_l P_{\pi_l} \right)}{logr} \tag{3 15b}$$

D_q is a nonincreasing function of q

There are other definitions of dimension that are useful in specific contexts The *topological dimension* is the dimension of the Euclidean space which the attractor resembles locally [70]. It is, by definition, an integer number The *pointwise dimension* is a local measure of the dimension [137], [39] It denotes, in the limit, the scaling of a measure of a box with its resolution around a local point on the attractor The *average* pointwise dimension is the corresponding global quantity given by an average of the pointwise dimension on the attractor Another dimension algorithm which gives an integer number even for fractal attractors is the *local intrinsic dimension* [47], [65] The basic idea is to compute the number of governing parameters in local regions The *average* local intrinsic dimension is the corresponding global average The *k-th nearest neighbour dimension* [39], [20] considers fixed mass boxes instead of fixed size boxes as in the case of generalized dimensions For strange attractors, there is a conjecture due to Kaplan and Yorke [78], that relates the dimension to Lyapunov exponents. Specifically, if $\lambda_1 \geq \lambda_2 \geq \quad \geq \lambda_d$ are the d Lyapunov exponents of the dynamical system, then the Lyapunov dimension D_L is given by

$$D_L = j + \frac{\sum_{k=1}^{j} \lambda_k}{|\lambda_{j+1}|} \tag{3 16}$$

where j is the largest index for which the sum over k is nonnegative The conjecture is that D_L is equal to the information dimension

Let us now consider the issue of numerical estimation of the generalized dimensions from time series data The theory of state space reconstruction, discussed in chapter 2, gives a method for reconstructing the attractor from a scalar observable of the dynamics Consider a time series x_i, $i = 1, \quad , N'$ obtained from the time evolution of a dynamical system f through the smooth scalar observable function h A time delay reconstruction in R^d produces a vector time series

$$x_i^d = [x_i \; x_{i+k} \quad x_{i+(d-1)k}]^T, \quad i = 1, \quad , N \tag{3 17}$$

for an integer delay k and embedding dimension d and $N = N' - (d-1)k$ If m is the dimension of the manifold (generally unknown *apriori*) on which the dynamics f evolves, then $d \geq m$ is a necessary condition and $d \geq 2m+1$ is a sufficient condition to ensure that the reconstructed vector time series is an *embedding* Therefore the dynamical invariants evaluated from the dynamical system through s_i, $i = 0, 1, \quad ,$ eq (3 1), or the reconstructed time series $\mathbf{x}_i^d$, $i = 0, 1, \quad ,$ eq (3 17), are the same This is particularly useful when a physical model of the dynamical system is not available

A direct method for the numerical estimation of generalized dimensions is to estimate the P_{π_l} in eq (3 12) by counting the fraction of the total number of trajectory points $\mathbf{x}_i^d$, $i = 1, \quad , N$ in partition π_l, $l = 1, \quad , M(r)$ However, a major disadvantage of this method is the requirement of large data length N which rapidly becomes impractical with increasing embedding dimension d [62] One way of reducing the effect of these limitations is to estimate the dimensions numerically from the generalized correlation sum [125], [50], [110] As a special case, the computation of correlation dimension D_2 from the correlation sum has become a most popular algorithm for dimension calculation

There is an important approximation that relates the generalized correlation sums to $\sum_{l=1}^{M(r)} P_{\pi_l}^q$ in eq (3 12) Let us consider this approximation and the subsequent relation of D_q to the generalized correlation sum Let $\mathcal{B}_{\mathbf{x}^d}(r, d)$ represent a ball of

radius r around the point $\mathbf{x}^d$ on the reconstructed trajectory of embedding dimension d Also, let

$$B_{\mathbf{x}^d}(r,d) = \frac{\mu[\mathcal{B}_{\mathbf{x}^d}(r,d)]}{\mu[\mathcal{A}]} = P_{\mathcal{B}_{\mathbf{x}^d}(r,d)} \tag{3 18}$$

For a point $\mathbf{x}_j^d$ on the reconstructed trajectory, $B_{\mathbf{x}_j^d}(r,d)$ can be estimated from

$$B_{\mathbf{x}_j^d}(r,d) = \frac{1}{N-1}\sum_{\substack{i=1\\ i\neq j}}^{N}\Theta(r - \|\mathbf{x}_i^d - \mathbf{x}_j^d\|) \tag{3 19}$$

where $\Theta(arg) = 1$ for $arg > 0$, and $\Theta(arg) = 0$, otherwise The choice of the metric in eq. (3 19) determines, for example, whether the ball $\mathcal{B}_{\mathbf{x}^d}(r,d)$ is a d-dimensional sphere (L_2 norm) or a cube (L_∞ norm) However, the dynamical invariants are immune to the choice of the metric. Let $\Pi = [\pi_1, \quad , \pi_{M(r)}]$ be a finite partition of side resolution r of the attractor A where $M(r)$ is the number of partitions and let $P_{\pi_l} = \frac{\mu(\pi_l)}{\mu(A)}$ be the probability of π_l, $l = 1, \quad , M(r)$ Also let $P_{\pi_{l(j)}}$ be the probability of that partition π_l which contains the jth point of the trajectory $\mathbf{x}_j^d$ If this partition contains $N_{\pi_{l(j)}}$ points of the trajectory, then the important approximation is

$$B_{\mathbf{x}_j^d}(r,d) \approx \frac{N_{\pi_{l(j)}}}{N} = P_{\pi_{l(j)}} \tag{3 20}$$

The idea behind this approximation is that most points in $\pi_{l(j)}$ will be within r of $\mathbf{x}_j^d$ and although some points that are farther away will be ignored, others that are close enough but in neighbouring partitions will be counted The error in approximation is only a factor of order unity and certainly not larger than the number of nearest neighbour boxes [50]

The generalized correlation sum is given by

$$C_q(r,d,N) = \frac{1}{N}\sum_{j=1}^{N}\left[B_{\mathbf{x}_j^d}(r,d)\right]^{q \quad 1} \tag{3 21}$$

The following derivation shows the approximation between the generalized correlation sum and the probability of partitions It proceeds by substituting $P_{\pi_{l(j)}}$ for $B_{x_j^d}(r,d)$ and replacing the sum over points on the reconstructed trajectory with a sum over the partitions and a sum over the trajectory points in each partition

$$\begin{aligned} C_q(r,d,N) &= \frac{1}{N}\sum_{j=1}^{N}\left[B_{x_j^d}(r,d)\right]^{q-1} \\ &\approx \frac{1}{N}\sum_{j=1}^{N}\left[P_{\pi_{l(j)}}\right]^{q-1} \\ &= \frac{1}{N}\sum_{l=1}^{M(r)}\sum_{j=1}^{N} I_{\pi_l}(x_j^d)\,[P_{\pi_l}]^{q-1} \\ &= \frac{1}{N}\sum_{l=1}^{M(r)} N_{\pi_l}\,[P_{\pi_l}]^{q-1} \\ &= \sum_{l=1}^{M(r)}[P_{\pi_l}]^q \end{aligned} \tag{3 22}$$

Comparing eqs (3.12) and (3 22), we have

$$D_q = \lim_{r\to 0}\frac{1}{q-1}\frac{\log\left[C_q(r,d,N)\right]}{\log r} \tag{3 23}$$

This formula is easiest to evaluate for D_2 As seen from eq (3 21), $C_2(r,d,N)$ is an arithmetic mean However, this method of evaluating D_q does not work very well for $q \leq 1$

3.4 Definitions of Dynamical Entropy and Relation with Generalized Correlation Sum

The generalized dimensions of an attractor are *static* invariants since they only utilize the time invariant measure of the attractor However, entropy is a *dynamic* invariant because it quantifies the increase in uncertainty or the loss of information about the state of a system with time *Metric entropy* asymptotically quantifies the average rate of loss of information about the state of a dynamical system as it evolves in time The idea of characterizing dynamics by an information rate is due to Kolmogorov

while Sinai gave a rigorous definition and proved that it makes sense Thus, metric entropy is also known as the *Kolmogorov* or *Kolmogorov-Sinai entropy*

If a dynamical system is observed by making measurements, then at the time instant of a measurement we know about the state of the system with a certain level of uncertainty depending on the accuracy of the measurement The level of uncertainty about the state of the dynamical system may remain the same upto the next measurement, i e there is no loss of information of the previous measurement as happens for completely predictable trajectories On the other hand, for chaotic trajectories, the level of uncertainty of the state of the system increases with time (loss of information) till the next measurement which reduces the uncertainty back to the original level

The metric entropy K for a dynamical system f as in eq (3 1) can be defined as follows [63]. Consider a trajectory s_i, $i = 1, \quad , N$ of the dynamical system f The state of the system is measured at intervals of time T Let the state space be partitioned by a finite partition $\Pi = [\pi_1, \pi_2, \quad \pi_{M(r)}]$ of side resolution r and number of partitions $M(r)$ Also, let $P_{i_1, \ ,i_n}$ be the joint probability that $s(t = T) = s_1$ is in some partition $\pi_k = i_1$, $s(t = 2T) = s_2$ is in partition i_2, and $s(t = nT) = s_n$ is in partition i_n Then, according to Shannon, the quantity

$$K(n) = -\sum_{i_1, \ ,i_n} P_{i_1, \ ,i_n} \log P_{i_1, \ ,i_n} \tag{3 24}$$

is the average information required to locate the system on a particular trajectory $i_1^*, \quad , i_n^*$ with precision r (provided only the *apriori* probabilities $P_{i_1, \ i_n}$ are known) Therefore, $K(n+1) - K(n)$ is the additional average information required to predict in which partition i_{n+1}^* will the system state s_{n+1} be, provided we know $i_1^*, \quad , i_n^*$ That is, $K(n+1) - K(n)$ measures the average loss of information about the system from time n to $n+1$

K is defined as the average rate of loss of information

$$K = \lim_{T\to 0} \lim_{r\to 0} \lim_{n\to\infty} \frac{1}{nT} \sum_{l=1}^{n} (K(l+1) - K(l)) \tag{3 25a}$$

$$= \lim_{T\to 0} \lim_{r\to 0} \lim_{n\to\infty} \frac{1}{nT} K(n) \tag{3 25b}$$

$$= \lim_{T\to 0} \lim_{r\to 0} \lim_{n\to\infty} \frac{1}{T} [K(n+1) - K(n)] \tag{3 25c}$$

For a more rigorous definition of metric entropy and a discussion on the importance of *generating* partitions in its computation, see [152], [31] The equality between eqs (3 25a) and (3 25b) is easy to establish. The equality between eqs (3 25a) and (3 25c) can be arrived at by noting that the former is a limit of the Cesaró sum of the sequence obtained from the latter Thus, it can be proved that the form of the expression in eq (3 25c) converges faster than the one in eq. (3 25a) [31]

Metric entropy can be estimated for both conservative and dissipative dynamical systems For completely predictable systems, $K = 0$, while for chaotic systems, $0 < K < \infty$ Ideal random behaviour is characterized by $K = \infty$ It is also important to note that K is inversely proportional to the average time over which the state of a dynamical system can be predicted This is because K quantifies the rate of loss of information about a system state Thus, the *predictability time* T_p is given by the proportionality relation [63], [128]

$$T_p \propto \frac{1}{K} \log \frac{1}{r} \tag{3 26}$$

where r is the precision with which the initial state of the system is located Beyond T_p, one can only make statistical predictions about the system state Thus, $T_p = \infty, 0 < T_p < \infty$ and $T_p = 0$ for completely predictable, chaotic and ideally random behaviours respectively There is also an interesting relation between the metric entropy and the spectrum of Lyapunov exponents given by

$$K \leq \sum (positive\ \lambda_i) \tag{3 27}$$

The equality which holds almost always for natural measures is known as the *Pesin identity* [36].

Just as we have generalized dimensions, there exists a set of infinite *order-q Renyi* or *generalized entropies* that characterize the space-time behaviour of dynamical systems These are defined as follows [50], [112], [116], [60]

$$K_q = -\lim_{T\to 0} \lim_{r\to 0} \lim_{n\to\infty} \frac{1}{nT(q-1)} \log \sum_{i_1, \ldots, i_n} P^q_{i_1, \ldots, i_n}, \quad q \neq 1 \tag{3 28a}$$

$$\lim_{q\to 1} K_q = K \tag{3 28b}$$

It can be seen that $K_q > K_{q'}$ for $q' > q$. Just like the correlation dimension D_2, the second order entropy K_2 is singled out because of its ease of calculation from a time series. Furthermore, since $K_2 \leq K$, we have $K_2 > 0$ for chaotic systems.

The method of generalized correlation sum can be extended to estimate the order-q entropies from time series data. Proceeding on similar lines as in the case of generalized dimension, one can easily show that

$$K_q = -\lim_{r\to 0}\lim_{n\to\infty} \frac{1}{nT(q-1)} \log\left[\tilde{C}_{q,n}(r,d,N)\right] \tag{3.29}$$

where

$$\tilde{C}_{q,n}(r,d,N) = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{1}{N-1}\sum_{\substack{i=1\\ i\neq j}}^{N}\Theta\left(r - \max_{l=0}^{n-1}\max_{m=0}^{d-1}\left|x_{i+l+m} - x_{j+l+m}\right|\right)\right]^{q-1} \tag{3.30a}$$

or

$$\tilde{C}_{q,n}(r,d,N) = \frac{1}{N}\sum_{j=1}^{N}\left\{\frac{1}{N-1}\sum_{\substack{i=1\\ i\neq j}}^{N}\Theta\left(r - \left[\sum_{l=0}^{n-1}\sum_{m=0}^{d-1}\left(x_{i+l+m} - x_{j+l+m}\right)^2\right]^{\frac{1}{2}}\right)\right\}^{q-1} \tag{3.30b}$$

for the L_∞ and the L_2 distance norms respectively. The reconstructed vector time series, given by eq. (3.17), is used with time delay $k = 1$ for simplicity of illustration above. The summation over m denotes the distance between two points $\mathbf{x}_i^d$ and $\mathbf{x}_j^d$ on the trajectory whereas the summation over l measures the distance between two sequence of points on the trajectory.

3.5 A Unified Approach to the Estimation of the Correlation Dimension and Second Order Entropy from Time Series

In this section we review a unified estimation procedure for generalized dimensions and entropies from a scalar time series observation of a dynamical system evolution. As mentioned before, the estimation of the correlation dimension D_2 and second order entropy K_2 from the correlation sum is a relatively easy exercise. The importance of

these invariants lies in their immediate relevance in a deterministic nonlinear state space modelling effort

Using the state space reconstruction theorems discussed in chapter 2 and recognizing that K_q is evaluated in the limit $n \to \infty$, we can replace this limit by $d \to \infty$ and the two summations over m and l for $\tilde{C}_{q,n}$ in eqs (3 30a) and (3 30b) by a single summation over m Thus [110]

$$\tilde{C}_q(r,d,N) = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{1}{N-1}\sum_{\substack{i=1\\ i\neq j}}^{N}\Theta\left(r - \|\mathbf{x}_i^d - \mathbf{x}_j^d\|\right)\right]^{q-1} \tag{3 31}$$

which is the same as $C_q(r,d,N)$ in eq (3 22) and

$$K_q = -\lim_{r\to 0}\lim_{d\to\infty}\frac{1}{dT(q-1)}\log\left[\tilde{C}_q(r,d,N)\right] \tag{3 32}$$

We now have $C_q(r,d,N) \sim r^{(q-1)D_q}$ from eq (3 23) and $C_q(r,d,N) \sim e^{(q-1)K_q}$ from eq (3 32) More specifically, observing the corresponding limits, we have

$$\lim_{r\to 0}\lim_{d\to\infty}\frac{1}{q-1}\log C_q(r,d,N) = D_q \log r - dTK_q \tag{3 33}$$

This yields, for $q = 2$, an expression for obtaining D_2 and K_2 from the correlation sum $C(r,d,N)$

$$\lim_{r\to 0}\lim_{d\to\infty}\log C(r,d,N) = D_2 \log r - dTK_2 \tag{3 34}$$

where

$$\begin{aligned} C(r,d,N) &= C_2(r,d,N) \\ &= \frac{1}{N(N-1)}\sum_{j=1}^{N}\sum_{\substack{i=1\\ i\neq j}}^{N}\Theta\left(r - \|\mathbf{x}_i^d - \mathbf{x}_j^d\|\right) \end{aligned} \tag{3 35}$$

and $\mathbf{x}_i^d$ is given by eq (3 17)

The practical implementation should now be clear [59], [24] We should obtain plots of $\log C(r,d,N)$ vs $\log r$ for increasing values of the embedding dimension d The estimation of D_2 from

$$D_2(d) = \lim_{r\to 0}\lim_{N\to\infty}\frac{\log C(r,d,N)}{\log r} \tag{3 36}$$

is inefficient because D_2 converges logarithmically slowly for $r \to 0$. Moreover, the correlation sum is dominated by additive noise at the $r \to 0$ limit which makes reasonable estimation difficult To overcome these problems, the usual procedure is to take a local slope at r away from zero or fit a straight line between two r values in which the graphs are approximately linear and obtain the slopes for increasing values of d.

$$D_2(d) = \frac{d \log C(r, d, N)}{d \log r} = \frac{dC(r, d, N)/dr}{C(r, d, N)/r} \tag{3 37}$$

or,

$$D_2(d) = \frac{\log C(r_2, d, N) - \log C(r_1, d, N)}{\log r_2 - \log r_1} \tag{3 38}$$

One can also do a least squares fit between r_1 and r_2 The value of $D_2(d)$ should converge with increasing values of d, which is then read off as the correlation dimension D_2

Similarly, the quantity

$$K_2(d) = \frac{1}{T} \log \left[\frac{C(r, d, N)}{C(r, d+1, N)} \right] \tag{3 39}$$

converges to K_2 for increasing values of d. Note that in eqs (3 37) and (3 39) we have suppressed the notation of the explicit dependence on the resolution r and data length N.

In the next section, we give the dimension estimation results from speech signals spoken in the form of phonemes by various speakers and from a simplified statistical model for a specific speech signal

3.6 Correlation Dimension Estimation for Speech Time Series

A dimension analysis of the vocal tract system has to take into account the time varying nature of the dynamics It is meaningful to estimate the dimension of a dynamical system only when the characteristics of the object (the attractor) do not

change in the time interval under investigation. Because of the physical nature of the vocal tract, we assume that its characteristics remain constant for some time We will do the dimension analysis of the vocal tract system based on speech signals When we compute the dimension from a block of speech data, we are actually doing so for a particular reconstructed trajectory The relationship between a reconstructed trajectory and the natural invariant measure of the attractor of the underlying dynamical system is only hypothesised

3.6.1 Numerical Computation from Speech Time Series

The speech signals used for the computation of dynamical invariants is as given in database 1 in Appendix B It consists of 57 consonants of the IPA spoken by 4 trained persons (3 males and 1 female) and 8 cardinal vowels spoken 4 times by one person (Daniel Jones). To approximate the probability measure by the correlation sum in eqs (3 22) and (3 35), it is necessary to approximate the $N \to \infty$ limit, i e , use as large a data length as possible Because of the extremely short duration of utterances of plosives, they are excluded from this analysis Thus, dimension analysis was done on $4 \times 44 + 4 \times 8 = 208$ phonemes as in the case for Lyapunov exponent analysis in chapter 2 and second order entropy analysis in section 3 8 The graphs of $\log C(r, d, N)$ vs $\log r$ are obtained for these phoneme time series The embedding dimension d was varied from 1 to 16 in each case and 20 log linear divisions of r were used to approximately cover the entire range of distances for the reconstructed trajectories L_2-norm was used to evaluate the distances used in the correlation sum

Figure 3 1 shows a typical graph of $\log C(r, d, N)$ vs $\log r$ for increasing values of d The slopes $D_2(d)$ are obtained from a linear fit in the approximately linear region of the plots Figure 3 2 shows the corresponding plot of $D_2(d)$ vs. d for $d = 1, \quad , 16$ For comparison, we have also plotted $D_2(d)$ vs d for a white Gaussian noise sequence of the same length The time series was generated using a computer based random number generation algorithm For true random behaviour, $D_2(d) = d$ Table 3 1 summarizes the results of dimension analysis performed on the 208 phonemes The correlation dimensions reported here are based on the values of the slope at $d = 16$, i e , $D_2 = D_2(16)$ The mean value of D_2 over the phoneme set is obtained as 3 74

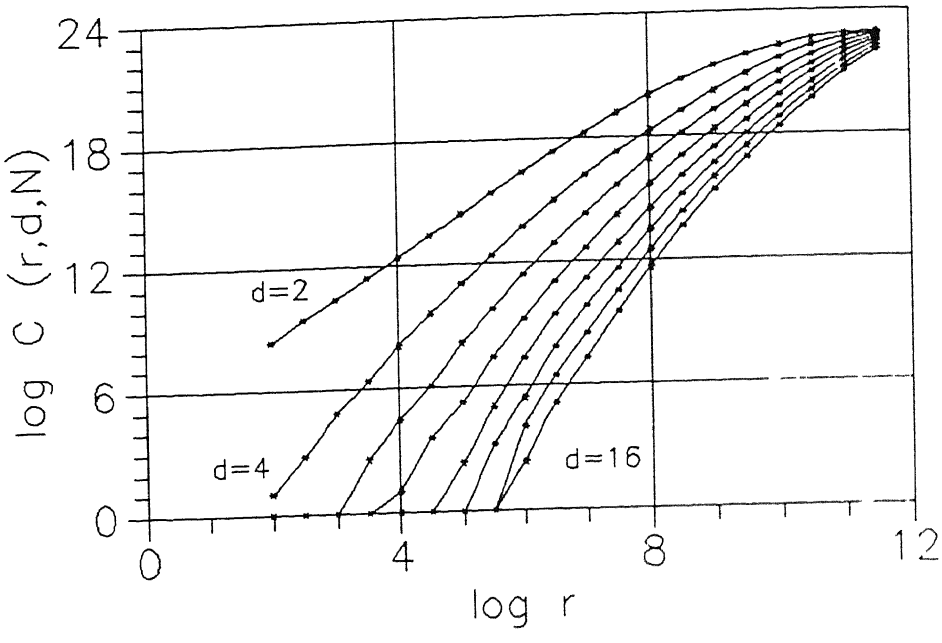


Fig. 3.1: Plots of $\log C(r,d,N)$ vs $\log r$ for $d = 2, 4, \ldots, 16$ and $N = 4000$ for cardinal vowel utterance /a/ DR-STV using 400 sample segments = 4.39 dB

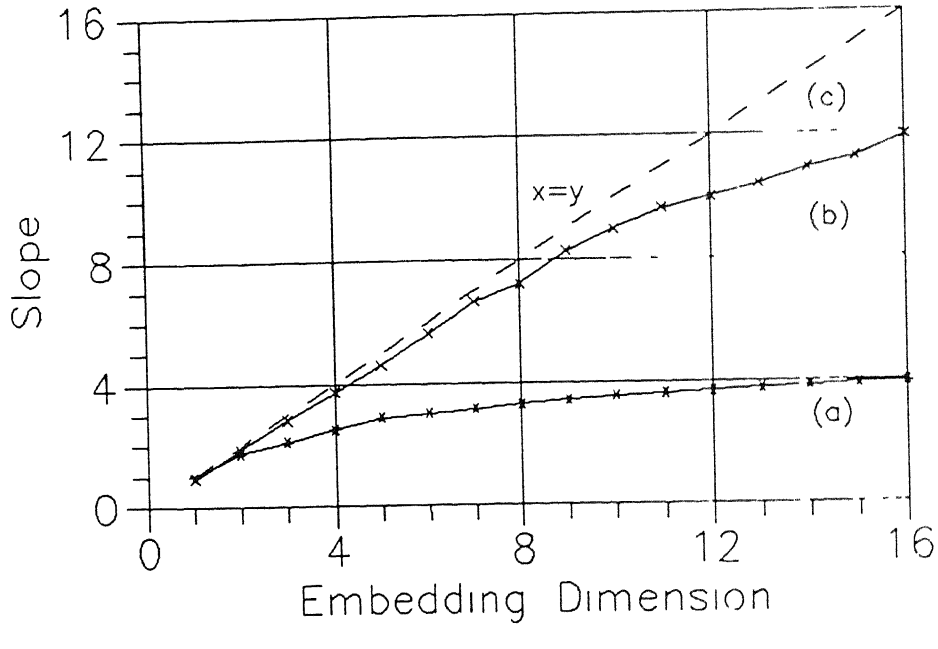


Fig. 3.2: Graphs for the computation of D_2 for (a) cardinal vowel utterance /a/ using the $\log C(r,d,N)$ vs. $\log r$ plots of Fig 3.1, and, (b) Gaussian white noise sequence of length $N = 4000$. Part (c) shows the ideal scaling of the slope for true stochastic time series

Phoneme Type	No of Phonemes	Sample [b]				Mean [c]
		1	2	3	4	
Vowel [a]	8	3 83 ± 0 51	3 62 ± 0 82	3 34 ± 0 82	3 36 ± 0 40	3 54 ± 0 20
Nasal	7	3 30 ± 0 52	2 37 ± 0 39	2 96 ± 0 25	1 90 ± 0 59	2 63 ± 0 54
Trill	2	4 53 ± 0 05	2 96 ± 0 91	4 40 ± 0 11	2 79 ± 0 73	3 67 ± 0 80
Tap or Flap	2	3 97 ± 0 50	2 96 ± 0 02	4 20 ± 0 10	2 75 ± 0 15	3 47 ± 0 62
Fricative	22	4 07 ± 1 21	3 96 ± 1 32	4 87 ± 0 91	4 04 ± 1 05	4 23 ± 0 37
Lateral Fric	2	5 26 ± 0 50	6 76 ± 1 16	5 55 ± 0 23	5 18 ± 0 69	5 69 ± 0 63
Approximant	5	3 96 ± 0 79	3 22 ± 1 11	3 45 ± 0 44	2 68 ± 1 28	3 33 ± 0 46
Lateral Approx	4	3 18 ± 0 40	4 18 ± 0 98	3 33 ± 0 60	2 04 ± 0 32	3 18 ± 0 76
Mean [d]		3 92	3 68	4 24	3 30	**3 74** [e]

Notes

a The 4 samples of Cardinal Vowels are by one speaker (Daniel Jones) The consonant samples are by 4 different speakers (3 males *samples 1-3* and 1 female *sample 4*)
b The error values are the s d over the phonemes in each group
c The error values are the s d over the mean dimension over the 4 samples
d The mean is for the 44 consonants spoken by 4 speakers
e The global mean includes the cardinal vowels

Table 3 1 The correlation dimension D_2 over 208 phonemes of speech database 1 summarized according to phoneme categories

The mean dimension values over 4 speakers for lateral fricatives and fricatives are greater than 4.0 and together account for the highest and second highest values respectively amongst the 8 families of phoneme types that were analysed. From this study, we conclude that speech is largely a *low dimensional* signal.

We qualify the above results further with the following observations.

1. **Data lengths used in the correlation sum:** The data length N used in the analysis of each of the 32 vowel samples was fixed at 4000. For the 176 consonant samples N varies from 1200 to 4800. The average length over the 208 phonemes is 2986. The data lengths were determined by two factors – (a) the duration of utterance of each phoneme, and, (b) the dynamic range of short time variance (DR-STV) [75]. Speech signal, which is highly nonstationary has a DR-STV of approximately 40dB. However, for the computation of dynamical invariants, the data should come from a stationary statistic (ideally, DR-STV=0dB). We fixed, somewhat arbitrarily, an upper limit for the DR-STV at 10dB for each phoneme time series. To compute the DR-STV, the total length is fixed as the length N of the data used in the computation, whereas 25 ms (i.e., 400 samples at 16kHz sampling rate) of speech data is used for each segment. The mean DR-STV for the 176 consonants analysed is 4.6 ± 2.8dB whereas that of the 32 vowels is 2.2 ± 1.9dB.

2. **Multiplicity of scales:** In some plots of $\log C(r, d, N)$ vs $\log r$, two distinct scaling regions are observed at different resolutions r. For example, see fig. 3.3. Although dimension is defined in the limit $r \to 0$, in real world we approximate by finding the dimension at a finite value of r. In this regard, we quote Mandelbrot [97]: "What is the dimension of a ball of yarn? From a great distance it is effectively a point and appears zero dimensional, on approach it becomes a three dimensional solid, moving closer, we discern the one dimensional threads, which then become three dimensional again, the threads are again composed of fibres, etc. These different scaling regimes would produce rather extreme oscillations in a numerical estimate of dimension. Typically, when we are computing dimension we are interested in a given scaling range, but it may be very difficult to discern." In our case, whenever two distinct scales are observed at different resolutions r, we have computed the dimension corresponding to the scaling at the lower range of the resolution.

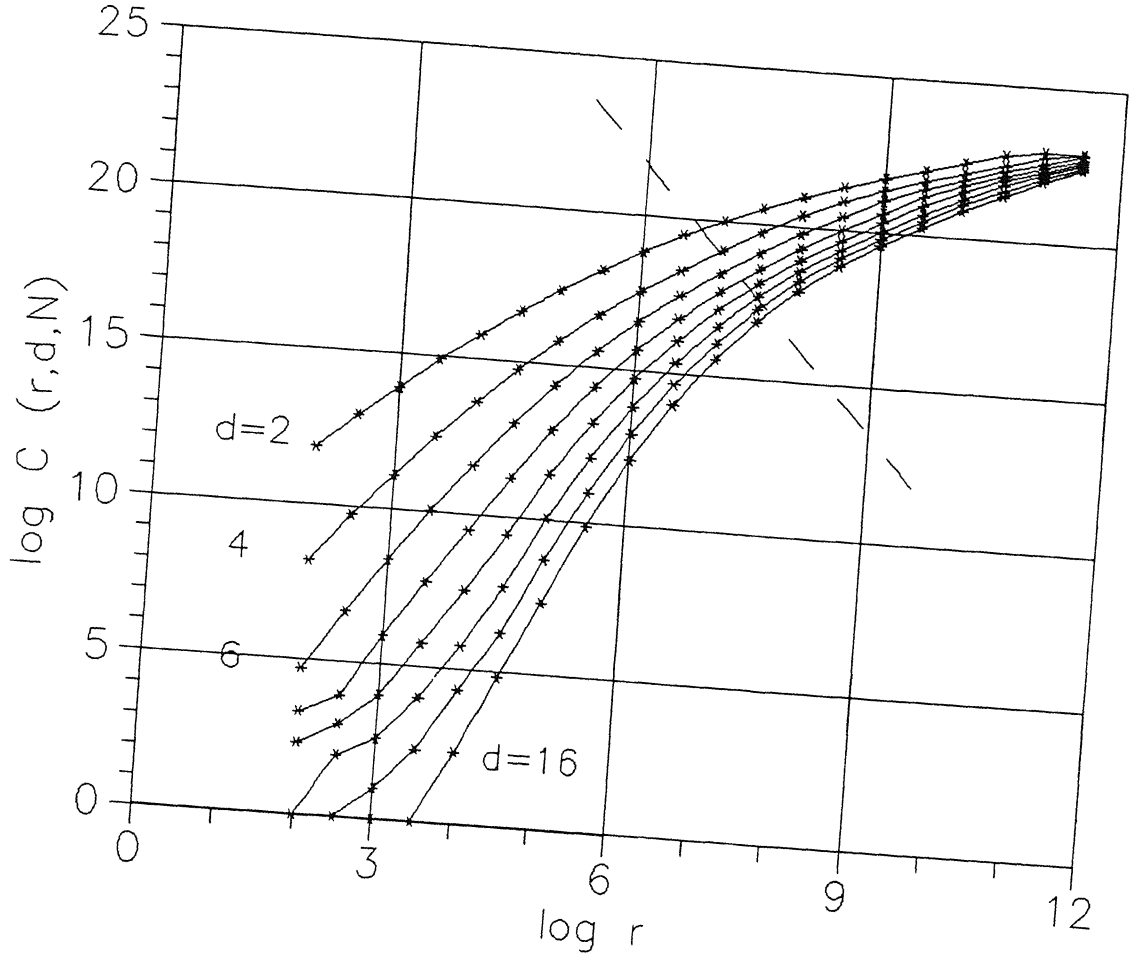


Fig 3.3: Plots to illustrate multiplicity of scales for cardinal vowel utterance /i/ Slopes on either side of the dividing line exhibit distinctly different scaling behaviour

3. Effect of autocorrelation on dimension estimation: The dimension results of Table 3 1 are computed from the modified correlation sum

$$C(r,d,N) = \frac{1}{(N-W)(N-W+1)} \sum_{j=1}^{N} \sum_{\substack{i=1 \\ |i-j|\geq W}}^{N} \Theta\left(r - \|\mathbf{x}_i^d - \mathbf{x}_j^d\|\right) \qquad (3\,40)$$

where W is chosen as 10 The purpose of the modification (of eq (3 35)) is to exclude those pairs of points $(\mathbf{x}_i^d, \mathbf{x}_j^d)$ from the correlation sum which are closeby in state space *only* because they are close in time It has been shown for specific cases [134], that this modification increases the scaling region in the $\log C(r,d,N)$ vs $\log r$ plot

If, however, we want to retain the effect of small time delay autocorrelation, then we can use the unmodified correlation sum of eq (3 35) No significant systematic difference in the dimension estimates is observed The mean value of D_2 over the 208 phonemes is obtained as 3 65 in this case

4. Comparison with filtered white noise: This test consists of comparing the correlation dimension of speech signals with that of linearly filtered white noise sequences having the same power spectrum The test data are created by taking the Fourier spectrum of each phoneme time series, randomizing the phase and inverting the transform We compared the correlation dimension of 13 phonemes (4 cardinal vowels, 1 nasal, 5 fricatives, 1 lateral fricative, 1 approximant and 1 lateral approximant) with that of linearly filtered white noise sequences having the same magnitude spectrum The corresponding mean values of the correlation dimension are $3\,32 \pm 1\,00$ and $4\,41 \pm 1\,04$ respectively Thus, there is a systematic increase in the dimension by 1 09 While it is possible to differentiate speech signals from white noise using spectral analysis and also dimension analysis (see fig 3 2), no *significant* difference in the dimension is observed between speech signals and linearly filtered white noise having the same magnitude spectrum This suggests that the low dimensionality of speech data is largely due to the second order characteristics of the data

It has been observed that random noise sequence with $\frac{1}{f^\alpha}$ power law spectra can exhibit finite correlation dimension when $\alpha > 1$, [105] Ideally, white noise

(coming from a stationary distribution) is characterized by infinite dimension This contradiction is resolved by observing that in the former case there is no underlying invariant measure [61] Therefore, one can only talk of finite dimension of *a realization* of random noise sequence with $\frac{1}{f^\alpha}$ power spectra, and not that of any underlying attractor or invariant measure Similarly, in light of the above result for speech signals, it is more appropriate to talk of the low dimensionality of speech signals and their corresponding reconstructed trajectories rather than linking them to any hypothesised attractor or invariant measure of the approximately stationary segments analysed

5 **Comparison with other results:** Preliminary results of the estimation of dynamical invariants were incorporated by us in [90], [92], [91] Other researchers have corroborated the result that speech is largely a low dimensional signal Dimension values between 3 and 5 for voiced speech are reported for a number of speakers in [140] Similarly, a correlation dimension of 3 3 over 30s of continuous speech is reported in [143], [144] In another study, [14], the correlation dimension from approximately 1*s* of vowel utterances [a , i , u] each by 3 male speakers were obtained between 1 2 and 1 9

3.6.2 Correlation Sum and Dimension from a simplified Statistical Model of Speech

In this subsection, we consider the estimation of the correlation sum and the computation of the correlation dimension from a statistical model of a vowel utterance The purpose of this exercise is to compare the result with the numerically estimated value of dimension for the same phoneme utterance The correlation sum given by eq (3 35) can be alternatively written as

$$C(r,d,N) = \frac{2}{N(N-1)} \sum_{n=1}^{N} \sum_{i=1}^{N-n} \Theta\left(r - \left\|\mathbf{x}_{i+n}^d - \mathbf{x}_i^d\right\|\right) \tag{3 41}$$

where $\Theta(\ )$ is the Heaviside function and $\mathbf{x}_i^d$ is given by eq (3 17) Assuming that the scalar observable x_i, $i = 1, \quad , N+d-1$, (i e , the speech time series) of the vocal tract system is a realization of a stationary random process, we can write

$$C(r,d,N) = \frac{2}{N(N-1)} \sum_{n=1}^{N} (N-n) P\left(\left\|\mathbf{x}_{i+n}^d - \mathbf{x}_i^d\right\| \leq r\right) \tag{3 42}$$

for some i (say, $i = 1$). Thus, we require the joint probability density function of the $2d$ random variables $x_1, \quad , x_{1+(d-1)k}, x_{n+1}, \quad , x_{n+1+(d-1)k}$ where k is the integer time delay

Various probability density functions (p d f 's) for speech signals have been proposed based on experiments Amongst them are a specialized form of the gamma density and the Laplacian density functions [75], [100] Short time p d f 's of speech segments are better described by the Gaussian p d f irrespective of whether the speech segment is voiced or unvoiced [75] We are here interested in the analysis of a specific phoneme which was chosen as the cardinal vowel /i/ This is because of the relative ease in formulating a statistical model for this phoneme utterance A joint Gaussian density function is chosen for the corresponding time series r_i, $i = 1, \quad , N'$ ($N' = N + d - 1$) Its autocorrelation function (ACF) is approximated by

$$ACF(n) = \begin{cases} \cos\left(\frac{2\pi n}{n_T}\right) \exp(-an), & 0 \leq n \leq n_{cut} \\ 0 & n > n_{cut} \end{cases} \tag{3 43}$$

where $n_T = 85, a = \frac{22}{16000}, n_{cut} = 1500$ Figure 3 4 shows the estimated ACF from a realization of the vowel utterance of length $N' = 4000$ and its approximation by eq (3 43) The covariance matrix can be found from eq (3 43), and without loss of generality, the mean vector for the $2d$ variables can be chosen as 0 Thus, the joint density function is completely specified.

Because of the relationship between the $2d$ random variables of x_i^d and $x_{i\ n}^d$ in eq (3 42), we can simplify to d random variables $y_1, y_2, \quad , y_d$ through

$$\begin{aligned} \mathbf{y}_n^d &= [y_1, \ y_2, \quad , \ y_d]^T \\ y_l &= x_{n+1+(l-1)k} - x_{1+(l-1)k}, \quad l = 1, \quad , d \end{aligned} \tag{3 44}$$

It is easy to show that $\mathbf{y}_n^d$ will also have a joint Gaussian density function

$$p(\mathbf{y}_n^d) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Delta_{\mathbf{y}}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{y}_n^d)^T \ C_{\mathbf{y}_n^d}^{-1} \quad (\mathbf{y}_n^d)\right] \tag{3 45}$$

where $C_{\mathbf{y}_n^d}$ is the $d \times d$ covariance matrix of $\mathbf{y}_n^d$ and $\Delta_{\mathbf{y}_n^d}$ is its determinant The ij^{th} element of $C_{\mathbf{y}_n^d}$ is given by

$$\begin{aligned} c_{ij} &= E\left[\left(x_{n+1+(i-1)k} - x_{1+(i-1)k}\right)\left(x_{n+1+(j-1)k} - x_{1+(j-1)k}\right)\right] \\ &= 2ACF\left[(i-j)k\right] - ACF\left[n + (i-j)k\right] - ACF\left[n + (j-i)k\right] \end{aligned} \tag{3 46}$$

For simplicity of illustration, we use the L_∞-norm for the metric in eq (3 42) Ideally, the choice of metric does not affect the dynamical invariants Thus,

$$P\left(\left\|\mathbf{x}_{i+n}^d - \mathbf{x}_i^d\right\| \le r\right) = P\left(\max_{1\le l\le d} |y_l| \le r\right) \tag{3 47}$$

For $N \to \infty$ and the approximation $\frac{n_{cut}}{N} \to 0$, we have

$$\begin{aligned} C(r,d) &= \lim_{\substack{N\to\infty \\ \frac{n_{cut}}{N}\to 0}} C(r,d,N) \\ &= \frac{2}{N(N-1)} \lim_{\substack{N\to\infty \\ \frac{n_{cut}}{N}\to 0}} \left[\sum_{n=1}^{n_{cut}} (N-n) P\left(\max_{1\le l\le d} |y_l| \le r\right) + \sum_{n=n_{cut}+1} (N-n) P\left(\max_{1\le l\le d} |y_l| \le r\right)\right] \\ &= \frac{2}{N(N-1)} \lim_{\substack{N\to\infty \\ \frac{n_{cut}}{N}\to 0}} \left[\sum_{n=n_{cut}+1} (N-n) P\left(\max_{1\le l\le d} |y_l| \le r\right)\right] \\ &= P\left(\max_{1\le l\le d} |y_l| \le r\right) \end{aligned} \tag{3 48}$$

where

$$P\left(\max_{1\le l\le d} |y_l| \le r\right) = \int_{y_d=-r}^{r} \cdots \int_{y_1=-r}^{r} p\left(\mathbf{y}_n^d\right) dy_1 \cdots dy_d \tag{3 49}$$

Note that for $n > n_{cut}$, $p(\mathbf{y}_n^d)$ is no longer a function of the delay n, since there is no signal correlation for $n > n_{cut}$ and we assume a Gaussian density function

An analytical computation of the correlation dimension from eqs (3 48) and (3.49) at resolution $r \to 0$ will give $D_2(d) = d$ However, we are here interested in the computation of $D_2(d)$ at a *finite* range of the resolution r which is relevant in a practical situation To obtain a graph of $\log C(r,d,N)$ vs $\log r$ we evaluate eq (3 49) at different values of r for increasing embedding dimension d Figure 3 5 shows the plots for $d = 1, \ldots, 7$ The scaling exponent at $d = 7$ is obtained as 4 07 using a linear fit. For comparison, the scaling component at $d = 7$ obtained from a time series realization of the phoneme /i/ of length $N = 4000$ is $2\,95 \pm 0\,17$ We conjecture that the difference in the two values can be partly attributed to the finite data length in the latter case However, this does not affect the nature of the conclusion drawn from the dimension results of section 3 6 1 viz speech signal is largely low dimensional in the resolution scale range in which the analysis was performed

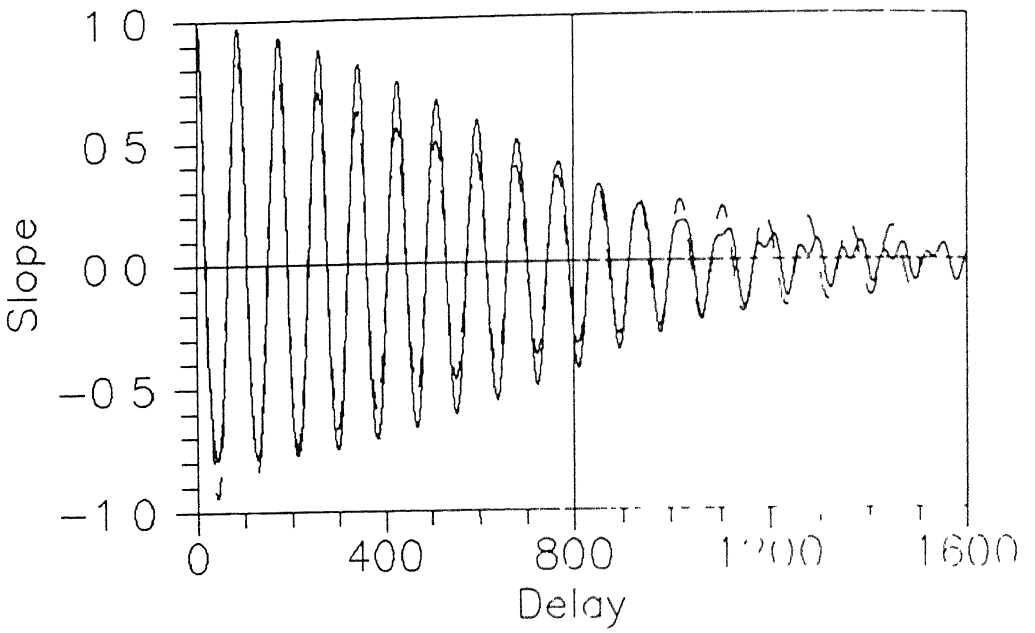


Fig. 3.4: The autocorrelation function for cardinal vowel utterance /i/ The graph shows the time series estimate from $N = 4000$ samples (solid line) and its approximation by eq (3 43) (dashed line)

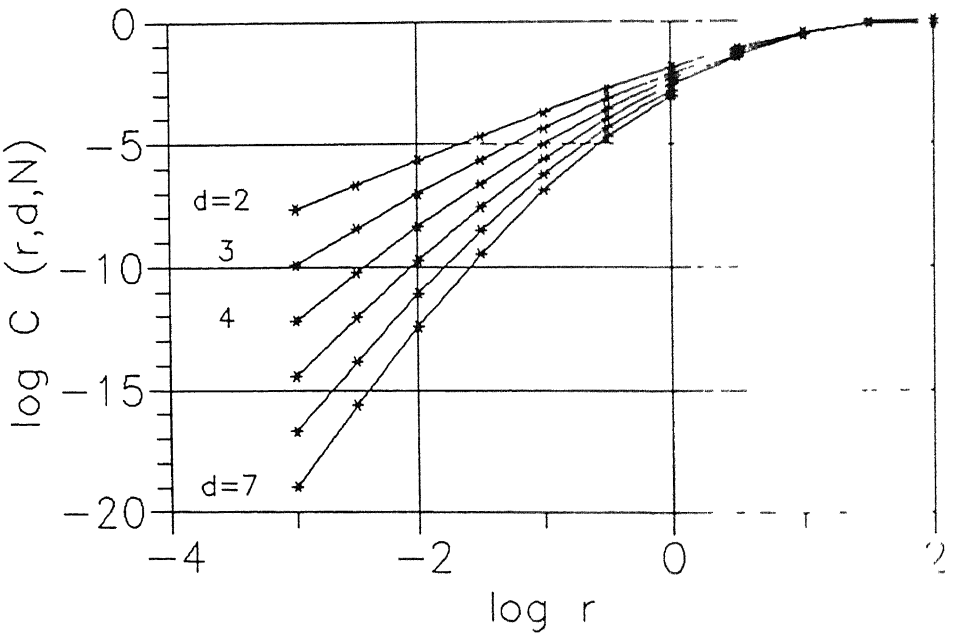


Fig. 3.5: Plots of $\log C(r, d, N)$ vs $\log r$ for $d = 2, 3, \ldots, 7$ from a numerical simulation of the theoretical estimate of the correlation sum in eqs (3 48) and (3 49)

In the next section we make some remarks about the limiting regions of the $\log C(r, d, N)$ vs $\log r$ plots from which the scaling exponents are computed We will also discuss about the various sources of error in dimension estimation.

3.7 Implementation Aspects of the Correlation Algorithm

More than one researcher has commented on the dimension estimation problem as being more of an art than a science In this regard, Theiler has remarked that experience indicates the need for more experience [137] The reason for such remarks is that the estimation of dimension from experimental time series data involves approximations and balances which are often qualitatively arrived at A large amount of research effort has been expended to make the estimation procedure more precise and to study mathematically the various sources of estimation error Because of finite data length and the presence of low amplitude additive noise, dimension estimation is done at an interpoint distance r away from 0 The procedure to do this is given in section 3 5 and in the following subsections we discuss the above mentioned issues further

3.7.1 Practical Remarks

A major advantage of the correlation algorithm over box counting algorithms for dimension estimation is that in the former case, one can probe down to distance scales of $O(N^{-\frac{2}{d}})$ i e , upto the smallest interpoint distance, whereas in the latter, one can only probe upto $O(N^{-\frac{1}{d}})$, i e , upto the average nearest neighbour distance, for data length N and embedding dimension d Another factor which restricts the scale length i e the extent of the abscissa in the $\log C(r, d, N)$ vs $\log r$ plots is the quantization of the data For 12-bit data, as in our case, the extent of the interpoint distance is over 13 octaves On the ordinate axis, the dynamic range of the correlation sum varies from $O(1)$ downto $O(\frac{2}{N^2})$ (For example, $\log_2(\frac{2}{N^2}) = -23$, for $N = 4096$, see fig 3 1) To facilitate the estimation of correlation dimension, it is desirable to divide the $\log C(r, d, N)$ vs $\log r$ plots into 4 possibly overlapping regions

1 At length scales of the order of the smallest interpoint distance the correlation sum scales as a collection of isolated points and hence is not suitable for dimension estimation

2 At small interpoint distance scales, the correlation sum may be inaccurate due to the presence of low amplitude additive noise in the data If the noise s d is σ, then the correlation sum at length scales below σ is expected to scale as the corresponding embedding dimension d [102]

3 Length scales above σ are ideal for estimating D_2

4 At length scales approaching the diameter of the attractor/state space extent of the trajectory, the correlation sum saturates This region again cannot be used for reliable dimension estimation

Another important practical issue is the efficient computation of the correlation sum A naive implementation would require the computation of $O(N^2)$ interpoint distances for each embedding dimension d considered, using data of length N Apart from eliminating the obvious redundant computations due to the nature of reconstruction vectors, eq (3 17), several algorithms have been proposed to reduce the computation effort A method to compute only the $O(N)$ shortest distances with $O(N \log N)$ effort is given in [135] Other algorithms use fast neighbour search routines [56] and multidimensional (k-d) trees [15] for this purpose

One popular method for reducing the computational complexity of the correlation sum is to choose few reference points $\mathbf{y}_j^d = \mathbf{x}_k^d$, $j = 1, \quad , N_{ref}$, $k \in (1, N)$ and use

$$C(r, d, N, N_{ref}) = \frac{1}{N_{ref} N} \sum_{j=1}^{N_{ref}} \sum_{i=1}^{N} \Theta \left(r - \left\| \mathbf{x}_i^d - \mathbf{y}_j^d \right\| \right) \tag{3 50}$$

where $\mathbf{x}_i^d$ is given by eq (3.17) and d is the embedding dimension Compare this with eq (3 35) It is, however, shown in [138] that such a scheme for reducing the computational complexity is not desirable because it decreases the precision of the estimate To optimize the statistical error with the computation time, it is necessary to take N_{ref} of the order of N, i e , the entire data length available

In the following subsection, we qualify the problem of estimating the correlation sum and dimension with a discussion of the various sources of error affecting their accuracy and precision. This is done in the next subsection

3.7.2 Sources of Error in Estimation

The various sources of error may affect the estimates in two fundamental ways – (a) It may cause the estimate to be *biased* This is variously referred to in dimension literature as *systematic error* or *accuracy* of the estimate There are wide variety of sources leading to this kind of error such as lacunarity, prefiltering, quantization, additive noise, edge effects etc. (b) The *statistical precision* or the standard deviation of the estimate from its mean value This arises from the finiteness of the data length available. The two kinds of errors usually work at cross purposes and they have to be balanced for optimal estimates

1. Lacunarity: The underlying assumption in estimating D_2 from $\log C(r, d, N)$ vs $\log r$ plots is that the correlation sum scales as a power of the interpoint distance The failure of this scaling to hold exactly is referred to as lacunarity This can be explicitly incorporated through a model $L(r)$ in the correlation sum as

$$C(r, d, N) = L(r)\ r^{D_2} \tag{3 51}$$

The middle-thirds Cantor set is a simple example which shows periodic lacunarity, i e , $L(r) = L(kr)$, $k = \frac{1}{3}$, and for all r Lacunarity has the effect of introducing intrinsic oscillations in the correlation sum If D_2 is estimated from a local slope, for example using eq (3 37), then a large systematic error may occur. Using eq (3 38) has the effect of reducing this systematic error particularly if several periods of the intrinsic oscillations are included in the scaling region (r_1, r_2) The reader is referred to [136] and its references for further details

2. Prefiltering: The phonemes of speech database 1 (Appendix B) were lowpass filtered at 7 5 kHz and sampled at 16 kHz The purpose of prefiltering is to attenuate high frequency noise in light of the fact that speech signals have negligible spectral content above this frequency level However, the process of filtering may lead to a systematic increase in the dimension estimate The process of filtering introduces

additional Lyapunov exponent(s) for the time series For example, an ideal lowpass filter with cutoff η introduces a new Lyapunov exponent $\lambda_f = -\eta$, the other exponents remaining unaffected [6] Consequently, the Lyapunov dimension, given by eq (3 16), remains unchanged as long as $\eta \geq |\lambda_{j+1}|$ Otherwise, an overestimate results

We have not computed the complete Lyapunov spectrum for the phonemes of the speech database However, the largest Lyapunov exponent for the 208 phonemes has been computed and the results summarized in chapter 2 The mean value is obtained as $2899s^{-1}$ In section 3.8, we present the results of second order entropy computation The mean value of K_2 is obtained as $8613\ 9s^{-1}$ Using these results along with Pesin's identity, eq (3 27), leads us to conjecture that the lowpass filtering at 7 5 kHz does not introduce a bias in the dimension estimates

The effect of convolution of a desired signal with a linear time invariant (LTI) system on the fractal dimension of an attractor is studied in [72] A sufficient condition is developed for the impulse response of the LTI system for which the dimension estimate is not affected Specifically, if the log magnitude of the largest pole radius of the LTI system is less than the *smallest* Lyapunov exponent, there will be no change in the dimension estimate due to the filtering Unfortunately, since this sufficiency test is in terms of the smallest Lyapunov exponent of a system, which is difficult to estimate accurately from time series observation of the dynamical system, this result is of little practical use

3. Quantization: This causes a clustering of points of the reconstructed trajectory on the vertices of a hypercube in the embedded space, where the side of the cube is equal to the least significant digit due to quantization Heuristically, this should lead to a systematic underestimate of the dimension because the clustering of the trajectory to a finite number of points reduces its complexity Upto first order, the estimate $\hat{D}_2$ of D_2 can be effectively modelled by

$$\hat{D}_2 = D_2 \left(1 - \frac{Kp}{\bar{r}}\right) \tag{3 52}$$

where K is a positive factor of order unity, p is one half of the least significant digit and $\bar{r} = (r_1\, r_2)^{\frac{1}{2}}$ (see eq (3 38)) [102] Assuming 12-bit quantization as in the

case of speech data, $r_1 = \frac{2^7}{2^{12}}$ and $r_2 = \frac{2^{10}}{2^{12}}$ at $d = 16$, we have $\frac{(D_2 - D_2)}{D_2} = -0\ 14\%$ For the estimated mean value of $\hat{D}_2 = 3\ 74$ for the speech data base, this represents a negligible systematic underestimate by 0 005 A standard procedure for reducing this systematic error is to use dithering

4 Additive noise: If the s d of noise is comparable to the length scale used for computing D_2, then the estimate $\hat{D}_2$ continues to increase as d is increased, resulting in an overestimate and the appearance of *approximate* convergence. It is demonstrated in [102] that such an overestimate due to additive random noise occurs for *all* values of σ An effective model for $\hat{D}_2$ is given by

$$\hat{D}_2 = D_2 \left[1 + K \left(\frac{\sigma}{\bar{r}} \right)^2 \right] \tag{3 53}$$

where K is a positive factor of order unity, σ is the s d of additive noise and $\bar{r} = (r_1\ r_2)^{\frac{1}{2}}$ For 12 bit data, assuming $r_1 = \frac{2^7}{2^{12}}$, $r_2 = \frac{2^{10}}{2^{12}}$ and σ to be equivalent to the lower 6 bits at $d = 16$, i e., $\sigma = 2^6$, we have $\frac{(D_2 - D_2)}{D_2} = 3\ 1\%$. From this, the estimated mean value of $\hat{D}_2 = 3\ 74$ for speech database 1, represents a systematic overestimate of 0 11 for the above parameters

5. Autocorrelated data: Autocorrelation of the time series may result in an anomaly in the correlation sum This issue and its remedy are discussed in section 3 6 1

6. Statistical Error and Data Length: We have so far considered sources of error which lead to a bias in the dimension estimate The *statistical error* or *precision* is solely an artifact of the finite data length used for obtaining the estimates The issue of statistical error in the estimation of correlation *sum* is addressed in [138] It is shown that the statistical error of the estimate scales generically as $O(\frac{1}{\sqrt{N}})$ as $N \to \infty$ There are exceptions which show $O(\frac{1}{N})$ scaling.

The problem of minimum data length requirement for reliable dimension estimates has also been widely investigated and various bounds reported To begin with, the correlation sum varies from $\frac{2}{N^2}$ to 1 for a data length of N samples If $R = \frac{r_2}{r_1}$ is the ratio of the interpoint distances over which the dimension is estimated, then

$$N_{min}^2 \geq R^d \tag{3 54}$$

is an absolute lower bound for N for a d-dimensional trajectory/attractor [137], [138]

By appealing to the edge effect at large scales and *nearest neighbour* separation at small scales, it is argued in [130] that to obtain an estimate within 5% of its true value, N is bounded by

$$N_{min} \geq 42^d \tag{3.55}$$

However this exceedingly high data requirement has been variously rejected (see, for example [61]) because of the wrong assumption that the number of neighbours required to resolve the correlation sum down to a specified interpoint distance r is the same as the density of points of the reconstructed trajectory in a hypercube of side resolution r in the appropriate embedding space This is only true for box counting algorithms such as those used for estimating the fractal dimension where the data requirements are much higher

In another study, [104], the incorrect assumption of the above bound is noted They use the combined influence of the edge effect at large scales and *characteristic spacing* between neighbours to arrive at a moderate bound

$$N_{min} \geq \frac{[2\,\Gamma(d/2+1)]^{\frac{1}{d}}}{(A \ln R)^{\frac{d+2}{d}}} \left[\frac{2(R-1)\,\Gamma\left((d+4)/2\right)}{\left(\Gamma\frac{1}{2}\right)^2\,\Gamma\left((d+3)/2\right)} \right]^{\frac{d}{d+2}} \tag{3.56}$$

Here, d and R are the same as in eq (3.54) and A is the maximum error allowed in the estimate

In a paper due to Ruelle [118], a fairly optimistic bound on the required data length N in terms of the correlation dimension d is given as

$$N_{min} > 10^{\frac{d}{2}} \tag{3.57}$$

Compare this with eq (3.54)

As an example, for an estimate of the maximum correlation dimension $d = 5.4$, N_{min} should be greater than 388, 2519 and 501 according to eqs (3.54), (3.56) and (3.57) respectively using $R = 8$ and $A = 0.1d$ (i.e., allowing a 10% error in the estimate using eq (3.56)). By comparison, the mean data length over 208 phonemes used for dimension estimation is 2986 It is yet not clear as to what constitutes a good criterion

for obtaining minimum data length requirements The differing requirements for N_{min} obtained with various criteria point to this

3.8 Second Order Dynamical Entropy for Speech Time Series

The speech database used for the computation of second order entropy K_2 is the same as used in the computation of correlation dimension and largest Lyapunov exponent. The metric entropy K is an upper bound of K_2 although the difference between the two values is usually negligible One generally computes K_2 instead of K because it is relatively easy to do so from the correlation sum Figure 3 6 shows a typical plot of $K_2(d)$ vs d for $d = 1,$, 15, eq (3 39) This plot is for the cardinal vowel utterance /a/ and is obtained from the $\log C(r, d, N)$ vs $\log r$ graph shown in fig 3 1 for the corresponding time series Table 3 2 summarizes the second order entropy analysis results for the 208 phoneme based time series of speech database 1 These are based on the values of $K_2(d)$ at $d = 15$ The mean value of K_2 over 208 phonemes is obtained as $8614 s^{-1}$ at 16 kHz sampling rate As in the case of correlation dimension, the mean value of K_2 over 4 speakers is the largest for lateral fricatives and the second largest for fricatives There is no further general trend to classify phonemes based on the second order/metric entropy However, for a state space model of the form of eq (3 1) and associated initial condition (i c) s_0 of the vocal tract dynamics or the speech signal, the mean value of K_2 gives an estimate of its predictability time T_p Thus, according to eq. (3 26), for an i c of 12 bit precision, $T_p \sim 1\,4$ ms In chapter 5, we will check the veracity of this result by prediction using a model which assumes very little about the data characteristics Alternatively, metric or second order entropy gives the average amount of information that should be optimally provided to a dynamical system at every iteration to maintain the level of precision of the trajectory at that of the initial condition

Pesin's identity (eq (3 27)), which relates the metric entropy to the sum of positive Lyapunov exponents of a dynamical system, allows us to compare the respective results for the case of speech signals The equality, which is expected to hold, is not violated as can be seen by comparing the second order entropy results

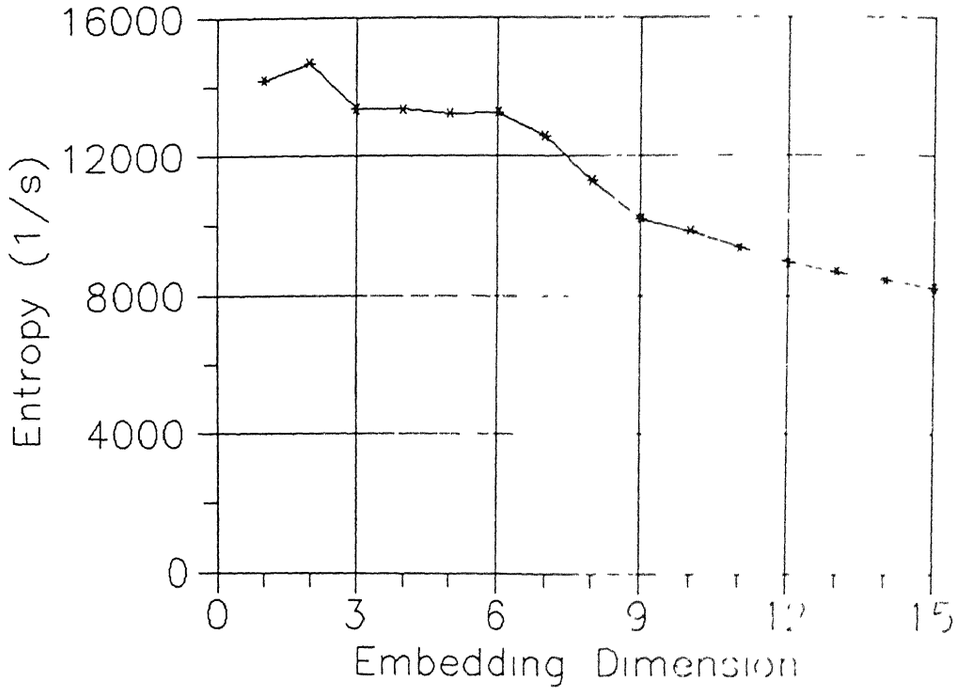


Fig. 3.6: Graph for the computation of the second order entropy for cardinal vowel utterance /a/ using the $\log C(r, d, N)$ vs $\log r$ plots of Fig 3.1

Phoneme Type	No of Phonemes	Sample [b]				Mean [c]
		1	2	3	4	
Vowel [a]	8	9893 ± 2580	8050 ± 1929	7690 ± 2263	8226 ± 1818	8453 ± 823
Nasal	7	7021 ± 1087	3798 ± 1016	7866 ± 1511	4730 ± 2169	5853 ± 1650
Trill	2	11929 ± 603	4768 ± 2736	12969 ± 1630	8252 ± 3539	9479 ± 3236
Tap or Flap	2	8135 ± 916	4744 ± 376	11478 ± 221	8805 ± 917	8290 ± 2399
Fricative	22	83487 ± 3775	7220 ± 5115	12660 ± 3544	9276 ± 4870	9497 ± 2030
Lateral Fric	2	12719 ± 2695	17519 ± 6937	14572 ± 144	11840 ± 48	14612 ± 2174
Approximant	5	8526 ± 1890	6450 ± 3878	7501 ± 1690	7352 ± 1933	7457 ± 737
Lateral Approx	4	5806 ± 1669	8625 ± 1892	7157 ± 1867	7659 ± 3147	7312 ± 1017
Mean [d]		8275 3	6959 9	10858 0	8460 7	**8613 9** [e]

Notes

a The 4 samples of Cardinal Vowels are by one speaker (Daniel Jones) The consonant samples are by 4 different speakers (3 males *samples 1-3* and 1 female *sample 4*)
b The error values are the s d over the phonemes in each group
c The error values are the s d over the mean entropy over the 4 samples
d The mean is for the 44 consonants spoken by 4 speakers
e The global mean includes the cardinal vowels

Table 3 2· The second order entropy K_2 over 208 phonemes of speech database 1 summarized according to phoneme categories

in Table 3.2 and the mean largest Lyapunov exponent values in Table 2.1. The positive values of the largest Lyapunov exponent and second order entropy both give evidence of the average divergence of nearby trajectories of speech signals in the reconstructed state space. This significant observation allows us to distinguish speech signals from strictly periodic or quasiperiodic behaviour.

Chapter 4

Polynomial Prediction of Speech

The dynamical analysis results of chapters 2 and 3 will form the basis of our investigation of some nonlinear prediction schemes for speech modelling and coding in a state space framework. The observation that speech is largely a low dimensional signal suggests that *few* state space variables will be needed to model it When a time series evolution is under observation, information is acquired at a rate of $I_a = -f_s \log_2 r$, where r is the resolution of the measurement and f_s is the sampling frequency However, the rate at which the underlying dynamics produces information is given by the metric entropy For speech database 1 (Appendix B), the second order entropy results of chapter 3 show that while information is acquired at a rate of 192 kb/s (12 bits/sample at 16 kHz), the information production rate is only of the order of 9 kb/s Over and above, if we allow for lossy coding of speech, which is normally the case, the required rate of transmission can be brought down further by a significant amount. (Note that the above figure of 9 kb/s is for 16 kHz sampling rate and not the usual 8 kHz for telephone grade speech) Also, if the estimation results of chapter 6, of the lower bound of the rate distortion function for speech sources assuming memory, are any indication of the possible limits to coding, then present day speech coders are still far off from them

Most present day speech coders in the low and medium bit rates, which reproduce natural sounding speech, are linear prediction based analysis - by - synthesis

coders Linear prediction has served speech coding purposes very well for decades and continues to do so The reasons for this are not difficult to find The theory of linear filtering is well developed, the superposition principle which is the cornerstone of linearity is often used to simplify coding complexity etc and the model of a source exciting a time varying linear filter can be seen to approximate the human speech production mechanism at certain levels

There have been isolated attempts in the recent past in the study of nonlinear representational forms for speech coding We have recorded them in the historical note of section 1 6 Most of the recent advances in speech coding, however, have come through manipulations of the excitation function rather than changes in the model form itself

We will study some nonlinear representation forms for predictive modelling of speech in a state space framework in chapters 4 and 5 This is a *deterministic* modelling framework as one for waveform coding must necessarily be The appeal of such a framework lies in the intuitive idea of the possibility of capturing apparent random or chaotic behaviour in compact model forms That this inverse problem is indeed addressable has been shown for specific examples in [33] There are two basic schemes for predictive state space modelling These are the global and local prediction schemes In a global prediction scheme, the function parameters are optimized over the entire state space whereas in a local prediction scheme the function is optimized over a local volume in state space where the prediction is to be done We consider these two basic approaches in chapters 4 and 5 respectively

The organization of this chapter is as follows In section 4 1, we review the salient features of the analysis - by - synthesis class of linear predictive coders We also discuss the basic structure and analysis steps of a CELP coding scheme Some model based analysis pointers to the presence of nonlinearities in the speech signal are given in section 4 2 Section 4 3 is concerned with the development of the analysis steps of polynomial prediction modelling of time series data A state space formulation of the modelling problem is also given in this section In section 4 4, we give results of our experiments on polynomial prediction of speech and its comparison with the traditional linear prediction and make some observations

4.1 Analysis - by - Synthesis Linear Prediction Coders

In this section, we describe the salient features of the analysis - by - synthesis class of linear prediction coders. Thereafter, we will study the code excited linear prediction (CELP) coder, which belongs to this class of coders Some of the early coders belonging to the analysis - by - synthesis class were the multipulse linear prediction coder (MPLPC) [4], the regular pulse linear prediction coder (RPLPC) [88] and the CELP coder [123]

The basic structure of an analysis - by - synthesis linear prediction coder is shown in fig 4 1 The name analysis - by - synthesis obtains from the fact that the coding parameters for successive blocks of speech are obtained by actually synthesising the reproduced speech in the coder itself Thus, the function of the decoder is emulated in the synthesis block of the coder also which is used to determine the optimal set of parameters meant for transmission The filter $\frac{1}{A(z)}$ models the short term correlations in the speech signal, i e , its spectral envelope It has the form

$$\frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{4 1}$$

where, a_i, $i = 1,$ $, p$ are the short term predictor coefficients and p is the order of the filter The typical values of p range from 8 to 16 assuming 8 kHz sampling rate, and the coefficients are determined from the speech signal in an open loop manner using linear prediction (LP) techniques [114]. Usually, the autocorrelation or stabilized covariance method [5] is used to determine the LP coefficients The filter coefficients themselves are computed from speech frames of 10 to 30 ms duration and adapted every frame For purposes of coding, the LP coefficients are normally converted to *line spectral pairs* (LSP), also known as *line spectral frequencies* (LSF), and quantized [131].

The filter $\frac{1}{B(z)}$ models the long term correlations in the speech signal corresponding to its spectral fine structure and has the general form

$$\frac{1}{B(z)} = \frac{1}{1 - \sum_{i=-q}^{r} b_i z^{-(D+i)}} \tag{4 2}$$

where b_i, $i = -q,$ $, r$ are the $(q + r + 1)$ long term prediction coefficients and D is usually the pitch delay The number of coefficients typically varies from 1 $(r = q = 0)$

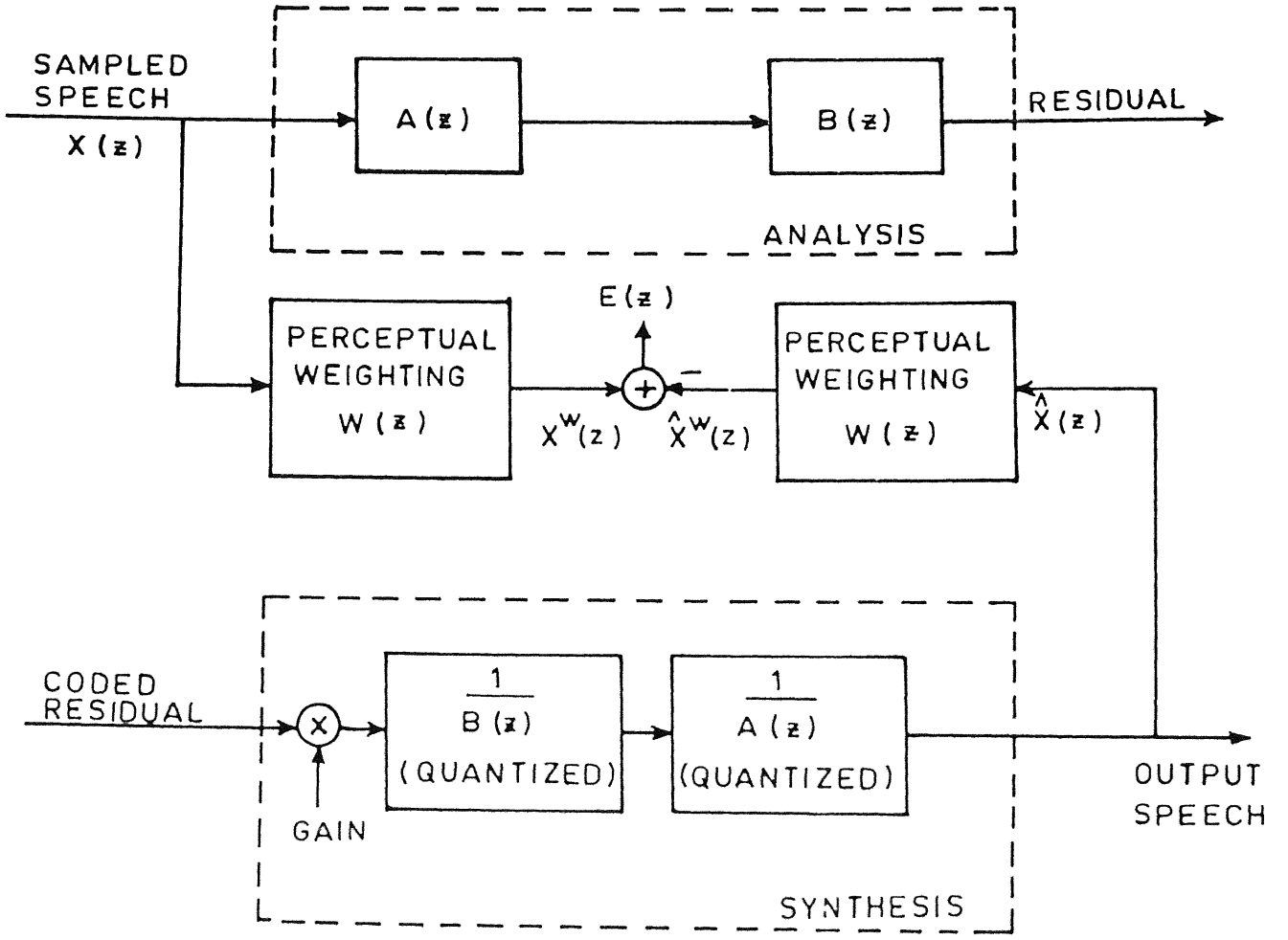


Fig 4.1· Block diagram of an analysis-by-synthesis linear prediction coder

to 3 ($r = q = 1$) The delay D and coefficients b_i, $i = -q,$ $, r$ are determined either from the speech signal or from the residual signal after removing the short term correlation as shown in the analysis block of fig 4 1 [115] The filter coefficients are usually adapted faster than the short term prediction coefficients at durations of 5 to 15 ms For a sampling rate of 8 kHz, a suitable range of the delay D is between 16 and 143 which corresponds to a pitch range between 56 and 500 Hz. The integral delay D can then be coded using 7 bits. The filter coefficients are generally scalar quantized for coding

The long term prediction filter is not always explicitly used in the coder. For example, in RPLPC and MPLPC, the excitation functions are often capable of modelling the pitch period redundancy In earlier versions of the CELP coder, a long term prediction filter was considered necessary However, nowadays its function is usually taken care of by an excitation function chosen from a so called *adaptive codebook* [81] This allows for an optimal closed loop search of the appropriate excitation function with relatively less complexity and also gives scope for incorporating *fractional* pitch delays [85] which improves the naturalness of reproduced speech.

The filter excitation function representing the coded residual can be chosen in a variety of ways Indeed a major portion of the research effort in the analysis - by - synthesis class of speech coders is given to the design of appropriate forms of excitation functions to serve various purposes In MPLPC, the residual is coded as a sequence of pulses located at *nonuniformly* spaced intervals The excitation analysis procedure has to determine both the amplitudes and positions of the pulses In RPLPC, the residual is coded as a set of *uniformly* spaced pulses The position of the first pulse within the excitation frame and the amplitude of the pulses are determined in the analysis procedure In a CELP coder, the excitation vector is chosen from a "shape" codebook and a scalar gain factor is chosen from a "gain" codebook in such a manner that a *weighted* distance between the original and reconstructed signal is minimized We will discuss the structure and operation of a CELP coder in more detail in section 4 1 1

The coder structure of fig 4 1 minimizes a weighted mean square error between the original and reproduced speech signals for each frame corresponding to the

duration of the excitation vector The weighting filter $W(z)$ incorporates a model of auditory perception based on the phenomenon of masking Its function is to reduce the perceived noise in the reproduced speech The effect of masking is to reduce the perception of one sound in the presence of another The human auditory mechanism has only a limited ability to detect small errors in the frequency bands where the speech signal has high energy (for example, in the formant regions) By minimizing the perceptually weighted mean square error, one makes use of the masking effect, in that the quantization noise is redistributed in relation to the speech power over different frequency bands A suitable weighting filter is given by, [87], [5]

$$W(z) = \frac{A(z)}{A(z/\gamma)} = \frac{1 - \sum_{i=1}^{p} a_i z^{-i}}{1 - \sum_{i=1}^{p} a_i \gamma^i z^{-i}}, \quad 0 < \gamma < 1 \tag{4.3}$$

where $A(z)$ is given by eq (4 1) and the parameter γ lies between 0 and 1 and controls the degree by which one wants to deemphasize the formant regions in the quantization error spectrum

For a more detailed review and a unified presentation of the basic coder forms of the analysis - by - synthesis class, the reader is referred to [87]

4.1.1 Structure and Analysis of a CELP Coder

In the analysis - by - synthesis class of linear prediction coders, the code excited linear prediction (CELP) coder is the most promising one in terms of performance parameters Consequently, significant attention has been given in the recent past to improve the CELP coder structure For a recent exhaustive survey of the advances made in this direction, see [51] In the following, we limit ourselves to a brief development of the essential features that constitute a unit analysis frame of a CELP coder

Figure 4 2 shows a structural block diagram of the CELP coder This is based on an alternative representation of the analysis - by - synthesis class of linear prediction coders compared to fig 4 1, and which is also frequently used (see, eg [86], [87])

The short term linear prediction analysis is done over one *analysis frame* of speech data Let us denote the sequence of data in an analysis frame by x_n, $n = 0, 1, \quad , N_f - 1$ The following analysis, however, is done over one *excitation frame* and is repetitive for each such frame Let us denote the sequence of data in an excitation frame by x_n, $n = 0, 1, \quad , N_e - 1$. The analysis frame length is chosen as an integral multiple of the excitation frame length Although the LP analysis is done over an analysis frame of speech data, the coefficients may be interpolated for each excitation frame in the LSP domain The LP analysis determines the coefficients of the two filters in the coder The function of the long term prediction filter, as shown in fig 4 1, is replaced by an excitation function chosen from an adaptive codebook and appropriately gain multiplied

For each excitation frame of speech data, x_n, $n = 0, 1, \quad , N_e - 1$, the perceptually weighted speech signal x_n^w, $n = 0, 1, \quad , N_e - 1$ is obtained from

$$x_n^w = x_n * h_n^w, \quad n = 0, 1, \quad , N_e - 1 \tag{4 4}$$

where, h_n^w is the truncated impulse response of $W(z)$ For obtaining the synthesised speech, the two filters $H(z) = \frac{1}{A(z)}$ and $W(z) = \frac{A(z)}{A(z/\gamma)}$ are combined into one filter given by

$$H^w(z) = \frac{1}{A(z/\gamma)} \tag{4 5}$$

The reproduced speech $\hat{x}_n^w$, $n = 0, 1, \quad , N_e - 1$ can be considered as a sum of the zero input response, z_n, $n = 0, 1, \quad , N_e - 1$ and the zero state response, y_n, $n = 0, 1, \quad , N_e - 1$ of the $H^w(z)$ filter corresponding to its excitation Thus,

$$\hat{x}_n^w = z_n + y_n, \quad n = 0, 1, \quad , N_e - 1 \tag{4 6}$$

The analysis - by - synthesis procedure must determine for the excitation frame, two indices corresponding to the excitation vectors from the adaptive and stochastic codebooks and two gain indices for the respective vectors by minimizing

$$E = \sum_{n=0}^{N_e-1} e_n^2 \tag{4 7}$$

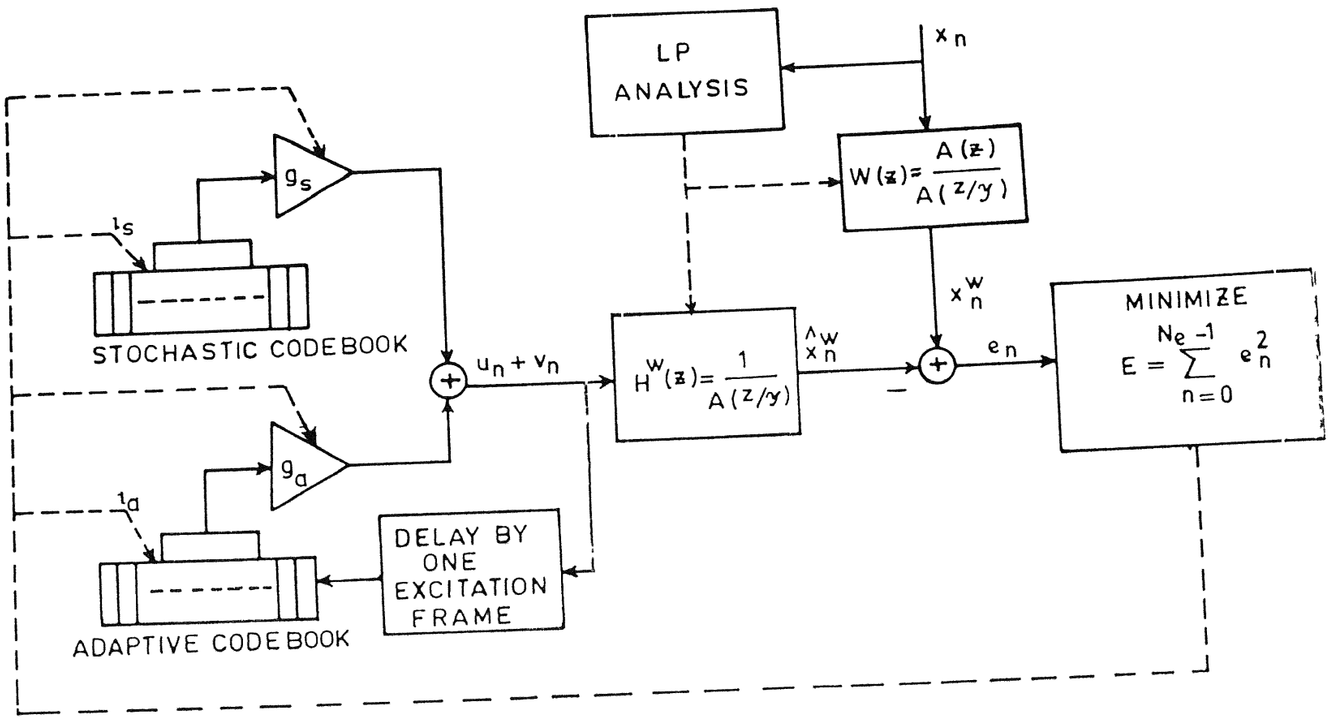


Fig 4.2 Structure of a CELP coder

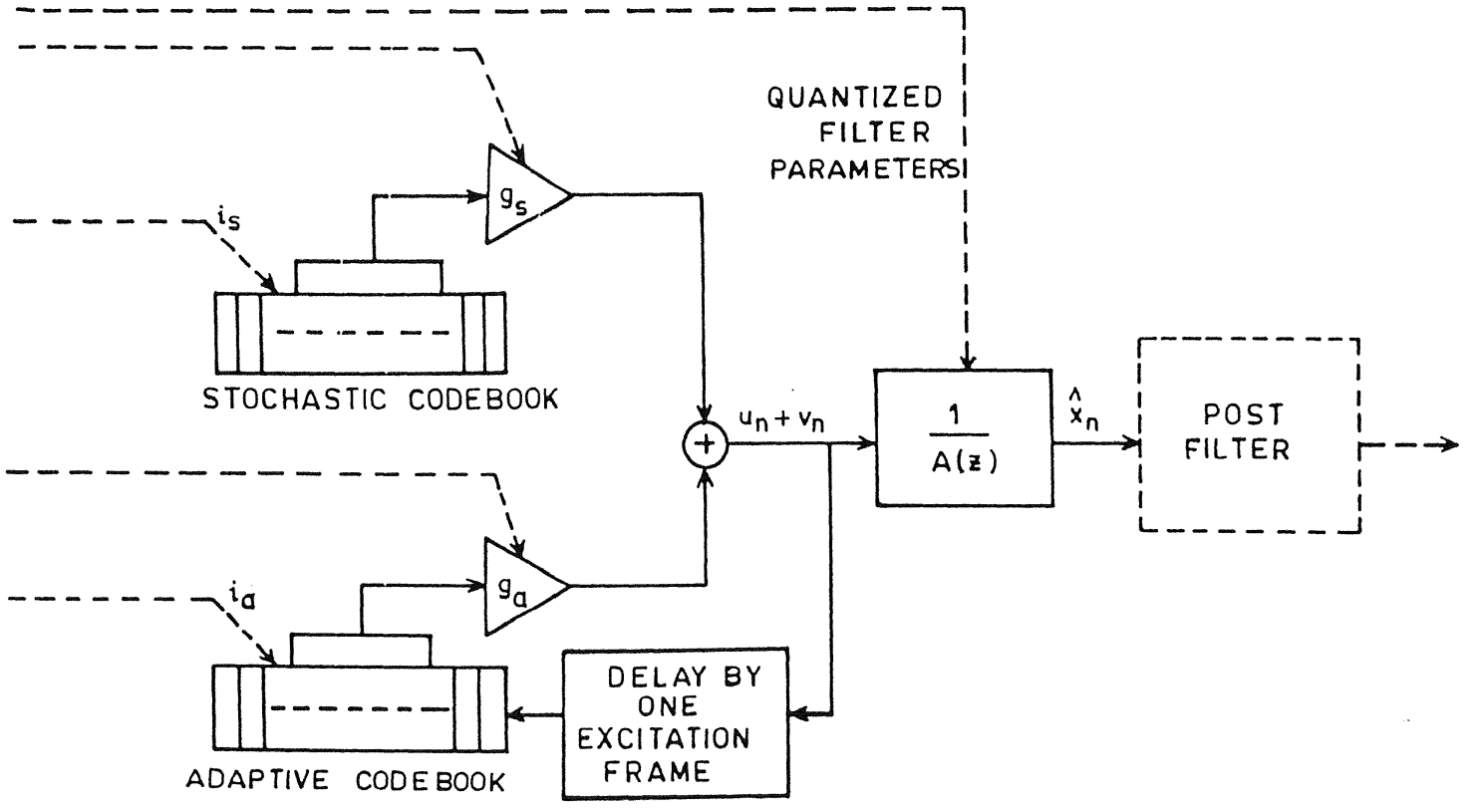


Fig. 4.3: Structure of a CELP decoder.

where

$$\begin{aligned} e_n &= x_n^w - \hat{x}_n^w \\ &= (x_n^w - z_n) - y_n \\ &= e_n^{(0)} - y_n, \quad n = 0, 1, \quad , N_e - 1 \end{aligned} \tag{4.8}$$

For notational convenience, we use vector representation for sequences of length N_e etc Thus, eq (4.7) can alternatively be written as

$$E = \mathbf{e}^{N_e T} \mathbf{e}^{N_e} \tag{4.9}$$

where

$$\mathbf{e}^{N_e} = [e_0 e_1 \quad e_{N_e - 1}]^T \tag{4.10}$$

We will henceforth drop the superscript since all the vectors are of length N_e

The minimization with respect to the adaptive codebook index and gain and the stochastic codebook index and gain is done sequentially in two steps In the first step of the minimization procedure, the indices corresponding to the adaptive codebook and its associated gain codebook are found In the second step, those of the stochastic codebook and its associated gain codebook are determined In the following, we discuss the two steps in the minimization procedure

STEP 1 – Adaptive Codebook Stage: The filter output due to the excitation from the adaptive codebook attempts to model the long term correlation structure of the speech signal The name "adaptive" codebook derives because it is updated after each excitation frame of speech is processed This way, the codebook tracks the varying long term correlation structure of speech Since the codebook update is based exclusively on the filter excitation vector of the previous frame which is also available to the decoder, no extra bits of information have to be transmitted to update the codebook on the decoder side [85]

Let $\mathbf{u}^{(i)}$ be a candidate vector given by

$$\mathbf{u}^{(i)} = g_a^{(i)} \mathbf{x}^{(i)}, \quad i = 1, \quad , N_a \tag{4.11}$$

where $\mathbf{x}^{(\imath)}$ is an excitation vector from the current adaptive codebook, $g_a^{(\imath)}$ is the corresponding optimal gain factor derived below and N_a is the size of the adaptive codebook In the first stage of minimization,

$$e_n^{(0)} = x_n^w - z_n, \quad n = 0, 1, \quad , N_e - 1 \tag{4 12a}$$

or,

$$\mathbf{e}^{(0)} = \mathbf{x}^w - \mathbf{z} \tag{4 12b}$$

where the vectors comprise of appropriate scalar entries according to eq. (4 10) Also,

$$\mathbf{y}^{(\imath)} = \mathbf{H}_w \mathbf{u}^{(\imath)}, \quad \imath = 1, \quad , N_a \tag{4 13}$$

is the scaled filtered codeword where

$$\mathbf{H}_w = \begin{bmatrix} h_0 & 0 & 0 & & 0 \\ h_1 & h_0 & 0 & & 0 \\ & & \cdot & \cdot & \\ & & \cdot & & \\ h_{N_e-1} & h_{N_e-2} & h_{N_e-3} & & h_0 \end{bmatrix} \tag{4 14}$$

is the $N_e \times N_e$ lower triangular matrix whose columns comprise the truncated impulse response of the $\frac{1}{A(z/\gamma)}$ filter The corresponding square error is given by

$$E^{(\imath)} = \mathbf{e}^{(\imath)^T} \mathbf{e}^{(\imath)} \tag{4 15a}$$

where

$$\mathbf{e}^{(\imath)} = \mathbf{e}^{(0)} - \mathbf{y}^{(\imath)} \tag{4 15b}$$

The optimal gain factor is obtained by minimizing $E^{(\imath)}$ with respect to $g_a^{(\imath)}$ and is given by

$$g_a^{(\imath)} = \frac{\mathbf{e}^{(0)^T} \mathbf{H}_w \mathbf{x}^{(\imath)}}{\mathbf{x}^{(\imath)^T} \mathbf{H}_w^T \mathbf{H}_w \mathbf{x}^{(\imath)}} \tag{4 16}$$

The problem then consists of finding the adaptive codebook index $\imath_a$ out of $\imath = 1, \quad , N_a$ which minimizes

$$E^{(\imath)} = -2\hat{g}_a^{(\imath)} \mathbf{e}^{(0)^T} \mathbf{H}_w \mathbf{x}^{(\imath)} + \left(\hat{g}_a^{(\imath)}\right)^2 \mathbf{x}^{(\imath)^T} \mathbf{H}_w^T \mathbf{H}_w \mathbf{x}^{(\imath)} \tag{4 17}$$

where $\hat{g}_a^{(i)}$ is the quantized value of $g_a^{(i)}$ obtained from a corresponding scalar codebook

STEP 2 – Fixed/Stochastic Codebook Stage: The name "fixed" codebook is to contrast it with the adaptive codebook since this is a prefixed codebook It is also referred to as a "stochastic" codebook because its candidate sequences are generally random in nature

Let $\mathbf{u} = \hat{g}_a^{(i_a)}\mathbf{x}^{(i_a)}$ be the first stage excitation vector The zero state response of $H^w(z)$ due to this input is given by

$$\mathbf{y}_1 = \mathbf{H}_w\mathbf{u} \tag{4.18}$$

Step 2 is similar to step 1 except in the codebook and the target vector $\mathbf{e}^{(0)}$ Here,

$$\mathbf{e}^{(0)} = \mathbf{x}^w - \mathbf{z} \quad \mathbf{H}_w\mathbf{u} \tag{4.19}$$

$$\mathbf{y}^{(i)} = \mathbf{H}_w\mathbf{v}^{(i)}, \quad i = 1, \quad , N_c \tag{4.20}$$

where N_c is the size of the stochastic codebook and

$$\mathbf{v}^{(i)} = g_s^{(i)}\mathbf{t}^{(i)} \tag{4.21}$$

$\mathbf{t}^{(i)}$ is an excitation vector from the codebook and $g_s^{(i)}$ is the corresponding optimal gain term obtained by minimizing

$$E^{(i)} = \mathbf{e}^{(i)^T}\mathbf{e}^{(i)} \tag{4.22}$$

where

$$\mathbf{e}^{(i)} = \mathbf{e}^{(0)} - \mathbf{y}^{(i)} \tag{4.23}$$

and $\mathbf{e}^{(0)}$ and $\mathbf{y}^{(i)}$ are given by eqs (4.19) and (4.20) respectively in this step The optimal value of $g_s^{(i)}$ is given by

$$g_s^{(i)} = \frac{\mathbf{e}^{(0)^T}\mathbf{H}_w\mathbf{t}^{(i)}}{\mathbf{t}^{(i)^T}\mathbf{H}_w^T\mathbf{H}_w\mathbf{t}^{(i)}} \tag{4.24}$$

The problem in this step then is to find the codebook index i, out of $i = 1, \quad , N_c$, which minimizes

$$E^{(i)} = -2\hat{g}_s^{(i)}\mathbf{e}^{(0)^T}\mathbf{H}_w\mathbf{t}^{(i)} + \left(\hat{g}_s^{(i)}\right)^2\mathbf{t}^{(i)^T}\mathbf{H}_w^T\mathbf{H}_w\mathbf{t}^{(i)} \tag{4.25}$$

Here $\hat{g}_s^{(i)}$ is the quantized value of $g_s^{(i)}$, (eq (4.24)), obtained from a corresponding scalar codebook

This completes the discussion of the analysis - by - synthesis steps for an excitation frame of speech data Figure 4.3 shows a block diagram representation of a CELP decoder. The information that is transmitted to the decoder once every excitation frame duration includes the adaptive codebook index, i_a, the corresponding gain index, the stochastic codebook index, i_s, and the corresponding gain index In addition, the quantized filter parameters are transmitted once every analysis frame.

Before closing this section, we note that the above CELP coding structure is only one of the common variants of the basic form. We have implemented a CELP coding scheme based on this structure for the purpose of comparison of the reproduced speech with that of a coding scheme to be proposed in chapter 5

4.2 Model based Indicators of Nonlinearities in Speech

In chapter 1, we made a general case for the study of nonlinear modelling schemes for speech based on certain observations of the speech production mechanism and the speech signal itself, limitations of a linear modelling scheme and recent advances made in the theory and practice of nonlinear processing methods In continuation of that theme, we present some experimental results based on speech models which give further indication of the need to study the performance of nonlinear models for speech This section can be looked upon as an interface between the nonlinear dynamical analysis work of chapters 2 and 3 and the nonlinear speech modelling and coding study to follow in the forthcoming sections and in chapter 5.

A. Dynamical Analysis of MPLPC Error Sequence

Some results of a correlation dimension analysis of the LP residual sequence are reported in [143], [144] In these studies, 10 ms segments of speech are used to adapt LP filter coefficients (model order = 12) and compute the dimension for 30s of the residual sequence obtained from successive speech segments It is reported that the scaling behaviour in the $\log C(r,d,N)$ vs $\log r$ plot (see, for example, sections

3.5 and 3.6) is less clear for the residual sequence compared to that of the speech signal. Also, the saturation of the correlation dimension is at a significantly higher value compared to that for speech signals. This is plausible because the analysed residual sequence is a *piecewise linear* (effectively nonlinear) transformation of the speech signal, and their corresponding correlation dimensions need not be simply related. However, if there is a deterministic explanation for speech signals, then its correlation dimension is theoretically the same as that of the residual sequence obtained by fitting a linear time series model (fixed coefficients) with finite number of lags [19]. This suggests that it is more appropriate to find the correlation dimension of the residual sequence obtained from individual speech segments and compare it with that of the corresponding speech signal. We will give some results of the dynamical analysis of reconstruction error sequences of an MPLPC scheme.

We have implemented a version of the MPLPC scheme in which iterative refinement of the predictor coefficients and the multipulse excitation is done [109]. The number of iterations was apriori fixed at 5 in which the parameters mostly converged. We did not use a long term predictor or the perceptual weighting filter. Other specifications are dictated by the requirement of large data length per speech segment, since we have to approximate the $N \to \infty$ limit in the correlation dimension algorithm. Each speech segment analysis contained 80 ms of individual phoneme articulations sampled at 8 kHz. The LP filter order p was fixed at 10 and 80 pulses were used to code the residual from each 80 ms speech segment. Thus, dynamical analysis was done on $N = 640$ samples of the reconstruction error sequence obtained from individual speech segments. Both the correlation dimension D_2 (eq. (3.36)) and the second order entropy K_2 (eq. (3.39)) were computed from 12 phoneme articulations comprising of the 8 cardinal vowels and one consonant each of types nasal, voiced and unvoiced fricative and approximant (Appendix B, database 1). Figure 4.4 shows a plot of $D_2(d)$ vs. d where d is the embedding dimension. Similarly, fig. 4.5 shows a plot of $K_2(d)$ vs. d for increasing values of d. The value of D_2 for the reconstruction error sequence of the above data set is obtained as 4.25 ± 1.33. Similarly, the value of K_2 is $7.731 \pm 1.897\ s^{-1}$. The error values indicate the standard deviation over the 12 samples.

For the purpose of comparison, the mean values of D_2 and K_2 for the speech signals of the same segments are obtained as 3.75 ± 0.79 and $7266 \pm 2424 s^{-1}$ respectively. We also evaluated D_2 and K_2 for white Gaussian noise sequences of the same data length. Ideally, $D_2(d) = d$ in this case. Similarly, K_2 should be infinite to indicate true randomness. Due to various approximations such as the use of finite data length, the computing technique and the method of generating the i.i.d. sequence itself (through pseudo-random number generators, which are chaotic maps) it is only possible to approximate the ideal results for D_2 and K_2. However, it is important to observe that the dynamical invariants are capable of distinguishing a random sequence from the reconstruction error sequence of an MPLPC scheme. This is evident from the comparative plots of figs. 4.4 and 4.5. In the case of the i.i.d. sequence, the value of $D_2(d)$ does not saturate with increasing d as seen from fig. 4.4. Similarly, the value of K_2 is significantly higher than that of the reconstruction error sequence (fig. 4.5).

Thus, the reconstruction error sequence is relatively low dimensional and has finite predictability. It can be argued that dynamical structure is still present in the error sequence which is amenable to state space modelling using a few number of variables. Alternatively, the speech signal itself may be modelled using a deterministic, nonlinear state space model.

In [143], [144] an experiment is described to determine the presence of a nonlinear deterministic component in the LP residual. A linear prediction analysis is carried out on the speech signal and the residual is replaced with Gaussian white noise of the same energy. This noise is then used to excite the LP filter to produce synthetic speech. The new signal is characterized by the property that its determinism is limited to the linear components – the best possible predictor of this signal is linear. By comparing the correlation dimension of this synthetic signal with that of the original speech signal, one can investigate the properties of the LP residual. The argument goes that if the correlation dimension of the noise excited LP filter is higher than that of the original speech signal, then one can conclude that the LP residual contains a nonlinear, deterministic component. The synthetic speech signal was found to have a correlation dimension of 3.9 compared to 2.9 of the original speech signal. From

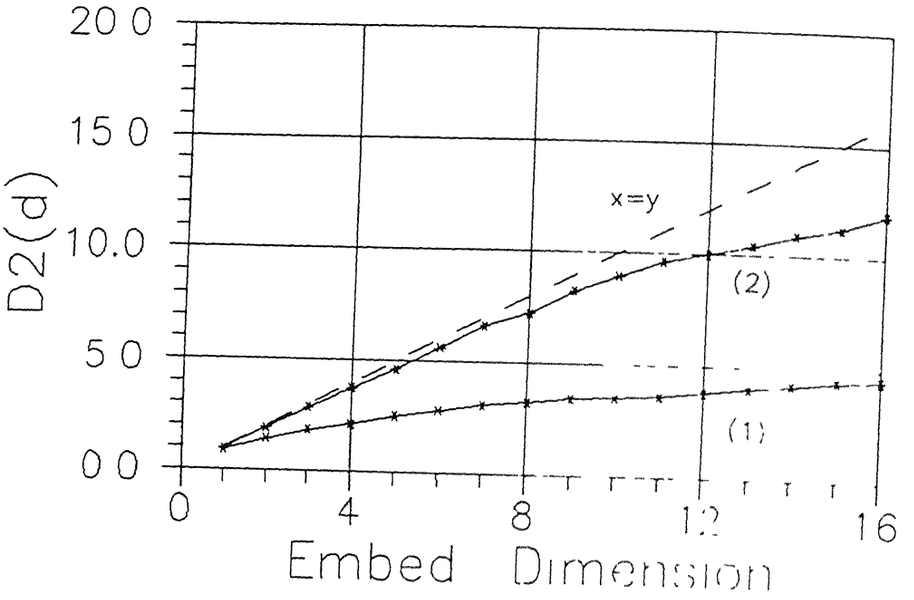


Fig 4.4: Graph for the computation of correlation dimension for (1) reconstruction error sequence obtained from an unvoiced fricative articulation /t/ using a MPLPC scheme, and, (2) white Gaussian noise sequence

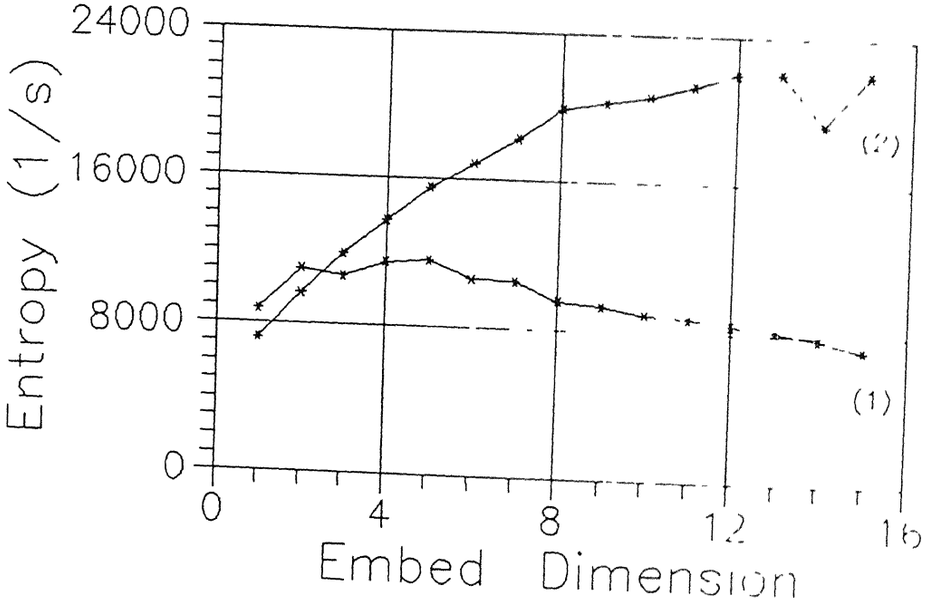


Fig. 4.5: Graph for the computation of the second order entropy for (1) reconstruction error sequence obtained from an unvoiced fricative articulation /t/ using a MPLPC scheme, and, (2) white Gaussian noise sequence

the higher value of the former, it is concluded that a suitable nonlinear predictor should be able to perform better than a linear predictor

B. Evidence from Polynomial Model of LP Residual

In an interesting experiment reported recently [139], direct evidence is given of the presence of short term quadratic nonlinearity in speech signals after removing *all* linear dependence with a linear predictor The steps of the experiment consist of the following

(i) Fit an optimal p^{th} order linear predictor to a frame of speech signal and obtain the residual sequence Call this the first stage residual. Compute the signal to prediction error ratio (SPER)

(ii) Fit an optimal p^{th} order linear predictor to the i^{th} stage residual sequence and obtain the $(i+1)^{th}$ stage residual sequence Compute the SPER Repeat (ii) until SPER $\sim$0 dB Let the final stage of LP analysis be the j^{th} stage When the SPER of the j^{th} LP analysis is 0 dB, all linear information within p samples will have been extracted from the original speech signal

(iii) Apply an optimal short term nonlinear predictor based on p samples to the j^{th} stage residual and get the $(j+1)^{th}$ stage residual Compare the j^{th} and $(j+1)^{th}$ stage residuals

The paper reports this comparison for a polynomial (quadratic) filter without the linear part Both voiced and unvoiced speech segments were studied The memory, p, of the linear as well as the nonlinear predictor is set to 10 samples Five successive LP analysis stages were sufficient to bring down the SPER to $\sim$ 0 dB It is observed that the nonlinear predictor provides an additional prediction gain over the LP stages (No comparative numbers are provided However, the residuals are graphically displayed) Further, pitch periodicity is almost completely removed with the short term nonlinear predictor which was still present even after the 5 stages of linear prediction

4.3 Analysis for Polynomial Prediction of Time Series

The choice of a polynomial representational form is usually one of the first out of the infinitely many possibilities This is because it is a simple extension of a linear

model and the optimal set of coefficients, in the sense of minimum m s e can be obtained by solving a set of simultaneous linear equations A polynomial predictor attempts to model the higher order correlations in time series If we look at the modelling problem as a purely waveform matching exercise, then the question of interest is whether a polynomial form gives better prediction gain compared to a linear model A deeper issue, which we have not investigated here, is that of the perceptual relevance of the new moment information that gets modelled

In predictive modelling, there are two time frames of interest One is the *prediction frame*, whose length d gives the lag upto which signal correlations are considered for modelling In linear prediction modelling, the model order p is equal to the prediction frame length d However, in a polynomial difference equation model, the total number of coefficients, p is given by $p = \binom{(d+l)}{d!l!}$, where l is the degree of the polynomial and d is the prediction frame length A disadvantage of a polynomial prediction scheme is that the number of model coefficients increases rapidly with d and l The second time frame of interest is the *analysis frame*, (length N_f) from which the optimal set of model coefficients are evaluated Broadly, the choice of this length is bounded on the lower side by the requirement of a stable solution for the coefficients and on the upper side by the data length available, or, as in the case of speech, by the length for which signal stationarity can be assumed

Let us consider the analysis steps leading to the determination of the optimal set of coefficients of the prediction model This development is analogous to the *covariance method* of linear prediction We assume an analysis frame length of data x_n, $n = 0, \quad , N_f - 1$ is available, over which the polynomial prediction model has to be optimized with respect to minimum m s e The covariance method requires knowledge of x_n, $n = -d, \quad , -1$ of the previous frame also, where d is the maximum prediction delay A polynomial difference equation is of the form

$$x_n = \sum_{m=1}^{p} a_m p_n^m + e_n \tag{4.26}$$

where

$$p_n^m \in \mathcal{P} = \{p_n^1, p_n^2, \quad , p_n^C\} \tag{4.27a}$$

and

$$p_n^m = x_{n-k_1}x_{n-k_2} \quad x_{n-k_i},\ 0 < i \le l,\ 1 \le k_1 \le d, \quad ,1 \le k_i \le d \tag{4 27b}$$

Here, d is the maximum prediction delay, l is the polynomial degree and $\mathcal{P}$ is the set of candidate polynomial terms The problem is to fit the equation

$$\hat{x}_n = \sum_{m=1}^{p} a_m p_n^m \tag{4 28}$$

to the data x_n, $n = 0, 1, \quad , N_f - 1$ in the analysis frame to minimize

$$E = \frac{1}{N_f} \sum_{n=0}^{N_f-1} e_n^2 \tag{4 29a}$$

where

$$e_n = x_n - \hat{x}_n, \quad n = 0, 1, \quad , N_f - 1 \tag{4 29b}$$

This leads to a set of p simultaneous linear equations

$$\sum_{m=1}^{p} a_m \left\langle p_n^i p_n^m \right\rangle = \left\langle x_n p_n^i \right\rangle, \quad i = 1, \quad , p \tag{4.30}$$

where $<\ >$ denotes a time average over N_f points Let us define

$$\phi^{i,m} = \left\langle p_n^i p_n^m \right\rangle, \quad i, m = 1, \quad , p \tag{4 31a}$$

$$\phi^{i,0} = \left\langle p_n^i x_n \right\rangle, \quad i = 1, \quad , p \tag{4 31b}$$

$$\phi^{0,0} = \left\langle x_n^2 \right\rangle \tag{4 31c}$$

Then, to obtain the optimal prediction model, one needs to solve the following for a_m, $m = 1, \quad , p$.

$$\begin{bmatrix} \phi^{1,1} & \phi^{1,2} & & \phi^{1,p} \\ \phi^{2,1} & \phi^{2,2} & & \phi^{2,p} \\ & & & \\ \phi^{p,1} & \phi^{p,2} & & \phi^{p,p} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \phi^{1,0} \\ \phi^{2,0} \\ \\ \phi^{p,0} \end{bmatrix} \tag{4 32a}$$

or

$$\Phi \mathbf{a} = \Psi \tag{4 32b}$$

The above equation can be solved with the Cholesky decomposition procedure This method allows us to compute the reductions in the average prediction error over the

analysis frame with the addition of successive terms in the model [114] It must be noted that we have not yet specified the order in which the terms are arranged in the model of eq. (4 26) This will determine the increase in average prediction gain as successive model terms are included from 1 to p We have principally considered two ordering schemes for the model terms

(1) In the first method, we exhaust all possible terms upto a certain time lag before considering terms which include signal dependence for greater lags While this does not completely specify the ordering of terms for a general degree l polynomial, the ordering for a quadratic polynomial, which we have studied for speech, is specifically given by $x_{n-1}, x_{n-1}^2, x_{n-2}, x_{n-1}x_{n-2}, x_{n-2}^2, x_n\ _3, x_n\ _1 r_n\ _3, r_n\ _2 r_n\ _3, r_n^2\ _3,$

(2) The second method, which is potentially more interesting, is based on orthogonal term search [82] Suppose that we are interested in considering nonlinear terms upto degree l and lag d The total number of possible candidate terms is given by $C = \frac{(d+l)!}{d!l!}$ of which only p terms are to be chosen such that the addition of each successive term in the polynomial model leads to the maximum reduction in prediction error over the analysis frame at that instant For this, Gram Schmidt orthogonalization is performed on the polynomial basis set (see eq (4 27)) This is combined with Cholesky decomposition for fast candidate search at each stage of term inclusion From eq (4 26), we have

$$x_n = \sum_{m\ 1}^{p} g_m w_n^m + e_n \tag{4 33}$$

where w_n^m is related to p_n^m by the Gram Schmidt orthogonalization procedure

$$\begin{aligned} p_n^1 &= w_n^1 \\ p_n^m &= w_n^m + \sum_{r\ 1}^{m-1} v_{mr} w_n^r, \quad m = 2, \quad , p \end{aligned} \tag{4 34}$$

Thus,

$$\begin{bmatrix} p_n^1 \\ p_n^2 \\ \cdot \\ p_n^p \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & & 0 \\ v_{21} & 1 & 0 & & 0 \\ v_{31} & v_{32} & 1 & & 0 \\ & & & & \\ v_{p1} & v_{p2} & v_{p3} & & 1 \end{bmatrix} \begin{bmatrix} w_n^1 \\ w_n^2 \\ \\ w_n^p \end{bmatrix} \tag{4 35a}$$

or

$$\mathbf{p} = \mathbf{V}\mathbf{w} \tag{4.35b}$$

The m.s.e. in terms of the original coefficients a_m, $m = 1, \ldots, p$ is given by

$$\begin{aligned} E &= \langle e_n^2 \rangle \\ &= \langle x_n^2 \rangle - \sum_{m=1}^{p} a_m \langle x_n p_n^m \rangle \\ &= \phi^{0,0} - \mathbf{a}^T \Psi \end{aligned} \tag{4.36}$$

where $\phi^{0,0}$ is given by eq. (4.31c) and $\mathbf{a}$ and Ψ are as given in eq. (4.32).

In terms of the orthogonal basis set $\mathbf{w}$, eq. (4.35), the m.s.e. is equivalently given by

$$E = \langle x_n^2 \rangle - \sum_{m=1}^{p} g_m^2 \left\langle (w_n^m)^2 \right\rangle \tag{4.37}$$

where g_m, $m = 1, \ldots, p$ are the coefficients given by eq. (4.33). Suppose that $a_r p_n^r$, $1 \leq r \leq p$ is the last polynomial term added to the model in eq. (4.26). The addition of this term reduces the m.s.e. by an amount

$$\Delta E_r = g_r^2 \left\langle (w_n^r)^2 \right\rangle \tag{4.38}$$

where

$$g_r = \frac{\langle x_n w_n^r \rangle}{\left\langle (w_n^r)^2 \right\rangle} \tag{4.39}$$

The polynomial term p_n^r should be chosen from the remaining candidate terms at the r^{th} stage such that ΔE_r is the maximum. A fast orthogonal search method to do this using Cholesky decomposition is given in [82]. This process is repeated until all the p terms are successively found.

We briefly look into the computational complexity of linear prediction and polynomial prediction schemes. There are two major sources contributing to the computational complexity in either scheme. These are the multiplications involved in the computation of the covariance matrix Φ and the solution of the p simultaneous equations to obtain the optimal set of coefficients. For the linear prediction scheme,

$O(N_f p)$ multiplications are required to obtain the Φ matrix and $O(p^3)$ multiplications are required to solve the simultaneous equations by Cholesky decomposition method which is used for the covariance method For the case of quadratic difference equations, $O(N_f d^3)$ multiplications are required to find the entries of the Φ matrix and $O(p^3)$ multiplications are required to solve the simultaneous equations by Cholesky decomposition for the first method of ordering of terms given above For the second method, a *fast* orthogonal search algorithm requires $O(pCN_f + p^3C)$ multiplications to obtain the optimal set of coefficients after the computation of the Φ matrix [82] In the above, N_f denotes the length of the analysis frame, p is the number of coefficients in the model, d is the maximum prediction lag considered and C is the total number of candidate terms from which the p polynomial terms are chosen

We have implemented the above two term selection procedures to study the performance of polynomial prediction of speech The results are elaborated in section 4 4

4.3.1 State Space Formulation of Polynomial Predictive Modelling of Time Series

The problem of polynomial prediction of scalar time series can be cast into a one-step prediction problem in d-dimensional state space, where d is the maximum time delay upto which signal correlations are considered in the polynomial model While such a reformulation may not offer additional insights in the present prediction problem, it will be immensely useful in the understanding of the local state prediction problem to be studied in chapter 5 We give this formulation here for the sake of completeness of discussion in terms of state space predictive modelling

Let us assume, as before, that time series data x_n, $n = -d,\quad ,N_f - 1$ is known Here, d and N_f are the prediction and analysis frame lengths respectively Reconstruct a d-dimensional vector time series $\mathbf{x}_n^d$, $n = -1,\quad ,N_f - 1$, where

$$\mathbf{x}_n^d = [x_{n-d+1} x_{n-d+2} \quad x_{n\ 1} x_n] \tag{4 40}$$

Then, a state space model of $\mathbf{x}_n^d$ is given by

$$\mathbf{x}_n^d = \mathbf{g}(\mathbf{x}_{n\ 1}^d) + \mathbf{e}_n^d \tag{4 41}$$

and

$$x_n = \mathbf{h}^T \mathbf{x}_n^d \quad n = 0, 1, \quad , N_f - 1 \tag{4 42}$$

where

$$\mathbf{e}_n^d = [0 \quad 0e_n]^T \tag{4 43}$$

$$h = [0 \quad 01]^T \tag{4 44}$$

$$g(\mathbf{x}_{n-1}^d) = \hat{\mathbf{x}}_n^d$$
$$= \left[x_{n-d+1} x_{n-d+2} \quad x_{n-1} f(\mathbf{x}_{n-1}^d)\right]^T \tag{4 45}$$

Let us denote

$$\hat{\mathbf{x}}_n = f(\mathbf{x}_{n-1}^d)$$
$$= \sum_{m=1}^{P} a_m p_n^m \quad n = 0, 1 \quad N_f - 1 \tag{4 46}$$

Then, e_n in eq (4 43) is given by

$$e_n = x_n - \hat{x}_n, \quad n = 0, 1 \quad N_f - 1 \tag{4 47}$$

$\hat{\mathbf{x}}_n^d$ is the *nominal* trajectory shadowing the original reconstructed trajectory $\mathbf{x}_n^d$ It differs from $\mathbf{x}_n^d$ only in the last coordinate To find the model coefficients, a_m, $m = 1, \quad , p$, eq (4 46), use any one of the following two equivalent m s e minimization criteria

(1) In terms of the norm of the vector difference, minimize

$$E = \frac{1}{N_f} \sum_{n=0}^{N_f - 1} \left\| \hat{\mathbf{x}}_n^d - \mathbf{x}_n^d \right\|^2 \tag{4.48}$$

where $\mathbf{x}_n^d$ and $\hat{\mathbf{x}}_n^d$ are given by eqs (4 40) and (4 45) respectively, and $\| \ \|$ refers to the L_2 norm

(2) In terms of the scalar difference between the original and reconstructed time series, minimize

$$E = \frac{1}{N_f} \sum_{n=0}^{N_f - 1} (x_n - \hat{x}_n)^2 \tag{4 49}$$

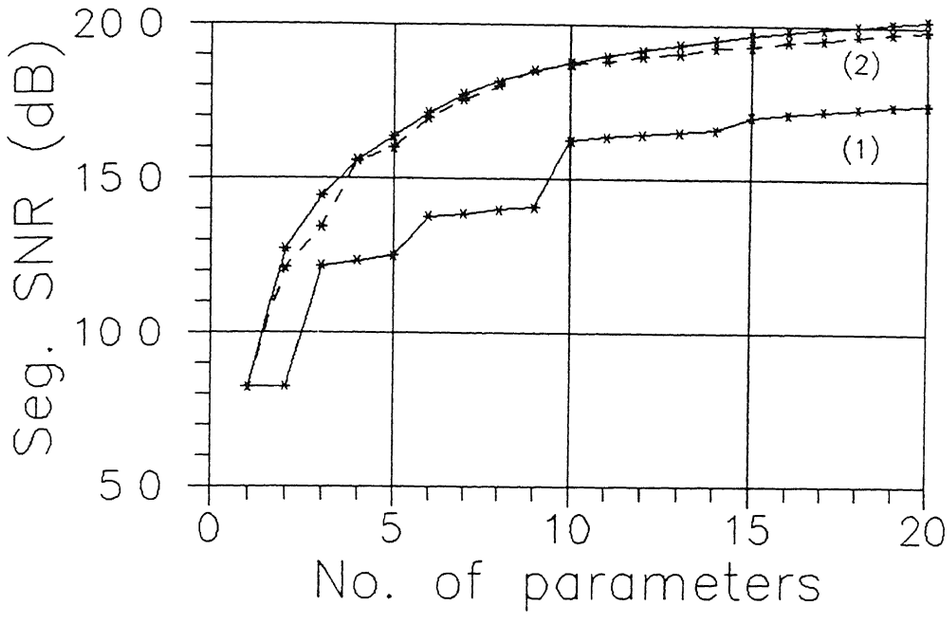
where x_n is the original time series and $\hat{x}_n$ is given by eq (4.46)

In the next section, we discuss the results of polynomial prediction of speech and make some observations

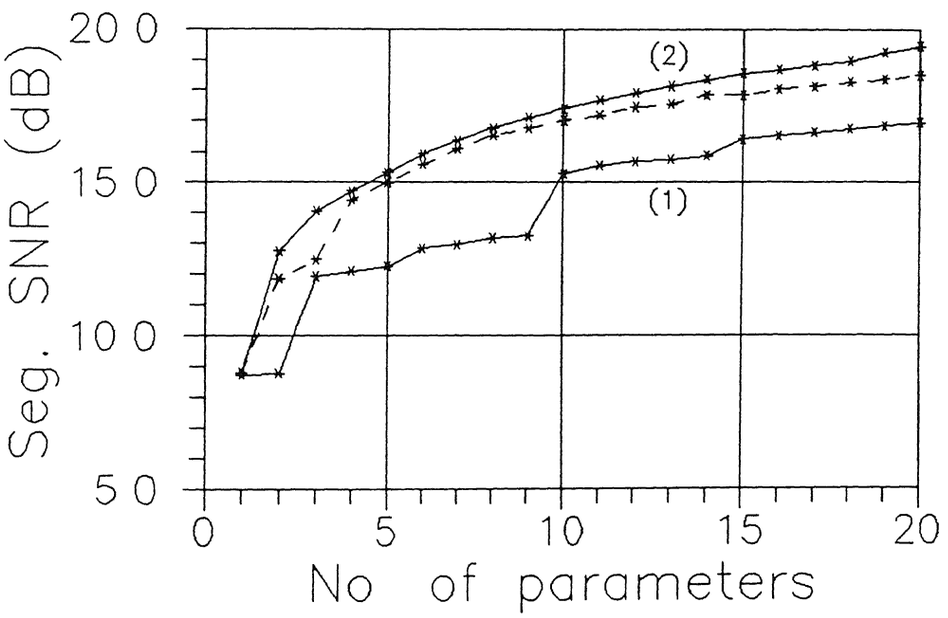
4.4 Results of Polynomial Prediction of Speech

We have implemented the polynomial prediction scheme and studied it for speech for the quadratic case, i.e., degree $l = 2$. Both the methods of ordering of model terms as discussed in section 4.3 are considered. Further, we have compared the prediction gains with that of a standard linear prediction scheme for the same number of coefficients in both models.

The phoneme specific sentence utterances of speech database 2 (Appendix B) were used as the basis of comparison. Each of the 4 sentences spoken by three males and three females were used for this study. The total duration of speech signal corresponding to each sentence spoken by the 6 speakers is as follows: sentence (a) Why were you away a year, Roy? – 14.25s, sentence (b) Nanny may know my meaning – 11.59s, sentence (c) His vicious father has seizures – 15.74s, and sentence (d) Which tea party did Baker go to? – 15.42s. Thus, a total of 57.0s of speech was used. Figures 4.6(a)–(d) show the segmental prediction gain (based on analysis frame length N_f) expressed as segmental SNR in dB for the two methods of term arrangement in the quadratic model for each of the four sentences respectively. The figures also show the corresponding segmental prediction gain for a linear prediction scheme (covariance method). Figure 4.6(e) shows the segmental prediction gains corresponding to the three cases averaged over all the four sentences. The analysis frame length, N_f was fixed at 160 samples corresponding to 20 ms of speech sampled at 8kHz. For the LP filter, the number of coefficients was varied from 1 to 20. For the first method of term arrangement in the quadratic model, the maximum delay d was fixed as 5. This allows a maximum number of 20 coefficients in the model without a constant term. The number of model terms were increased from 1 to 20 according to the term ordering strategy outlined in section 4.3. For the second method of term arrangement corresponding to orthogonal selection, the candidate set included all *linear* terms upto a delay of 10 samples and all *quadratic* terms upto


20 0
15 0
10 0
5 0
Seg. SNR (dB)
0
5
10
15
20
No. of parameters
(1)
(2)

(a)


20 0
15 0
10 0
5 0
Seg. SNR (dB)
0
5
10
15
20
No of parameters
(1)
(2)

(b)

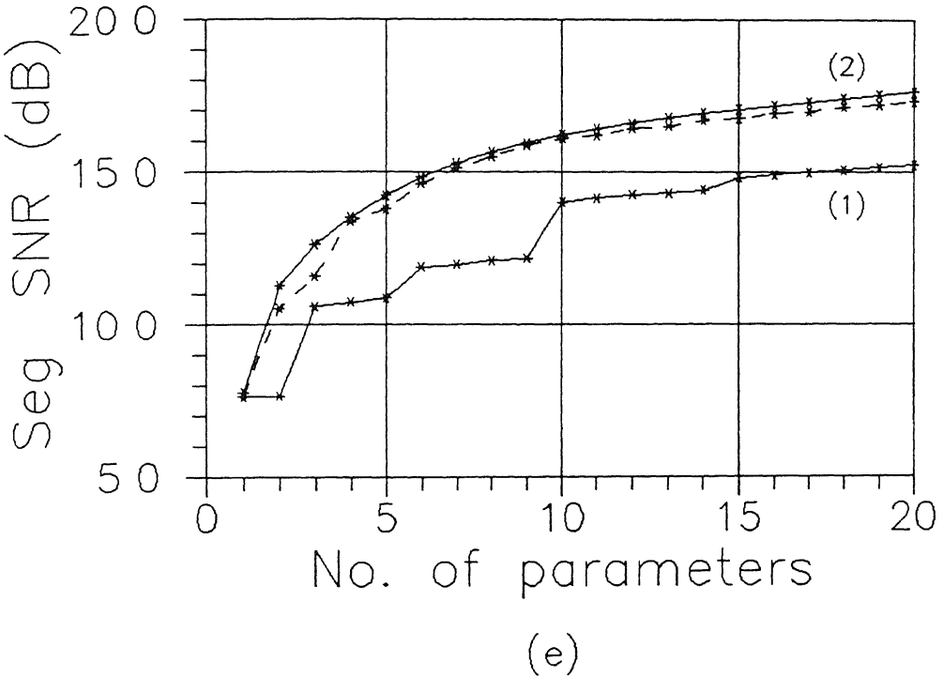


Fig. 4.6: The segmental prediction gain expressed as SNR in dB as a function of the number of model coefficients for three types of predictors (i) Linear predictor shown by dashed line, (ii) Quadratic predictor using first method of model term selection, denoted by (1), and (iii) Quadratic predictor using second method based on orthogonal term selection, denoted by (2) Parts (a)–(d) show the respective graphs for the 4 sentences (a)–(d) respectively (see text) spoken by 3 males and 3 females Part (e) shows the 3 graphs averaged over all the 4 sentences spoken by the 6 speakers

a delay of 6 samples Thus, from a total of $10 + 21 = 31$ candidate terms, 20 terms were successively selected for each analysis frame according to the orthogonal selection procedure of method 2 and the corresponding increase in prediction gain was recorded with the addition of each successive term

In context of the above, the following observations are noteworthy

(1) The quadratic prediction scheme, based on the first method of term arrangement, does not perform as well as the LP scheme in terms of segmental prediction gain (SNR) This can be seen from the graphs of figs 4 6(a)–(e) For 10 coefficients per model, the quadratic predictor gives 3 09 dB *less* segmental prediction gain compared to the linear predictor. At 20 coefficients per model, this difference is 2 08 dB

(2) It must be noted that while the linear predictor availed of the correlations upto a time delay of p samples, the quadratic predictor is based on signal dependencies upto a lag of d samples where $d < p$ For example, for $p = 9$, $d = 3$ and for $p = 20$, $d = 5$ In view of the poor prediction performance of the quadratic predictor based on the first method of term arrangement, it may be worthwhile to first remove the linear dependence in the signal and then apply the nonlinear predictor to the residual signal Based on this intuitive idea, we studied the performance of the following two stage predictors In the first stage, an optimal 10^{th} order LP filter is applied to the speech signal per analysis frame to give the first stage residual In the second stage, a quadratic predictor (1 to 20 coefficients based on a maximum prediction delay, d, of 5 samples) was optimally determined from the first stage residual for each analysis frame, to get further prediction gain We compared the prediction performance with a LP filter (number of coefficients varying from 1 to 20) at the second stage The additional segmental prediction gain at the second stage obtained over the entire speech database with the quadratic predictor is 2 07 dB for 10 coefficients and 2 83 dB for 20 coefficients compared to 0 54 dB and 1 17 dB for 10 and 20 coefficients respectively with the LP

(3) The second method of terms arrangement of the quadratic predictor gives a modest improvement in the segmental prediction gain compared to the LP case for the same number of coefficients This can be seen from the comparative graphs of the

overall segmental SNR in fig 4 6(e) and for sentences (a), (b) and (c) in figs 4 6(a)-(c) respectively However, for sentence (d), the LP gives slightly better segmental SNR for number of model terms exceeding 8 For 10 coefficients per model, the quadratic predictor gives 0 11 dB improvement in overall segmental prediction gain compared to LP At 20 coefficients per model, the improvement is 0 33 dB (fig 4 6(e)).

(4) For the orthogonal term selection procedure of the quadratic model, we recorded the frequency at which each term was selected at a particular position, for all positions in the model It was found that the terms x_{n-1}, x_{n-2}, x_{n-3} and x_{n-4} were selected more often at the 1^{st}, 2^{nd}, $3^{rd} and 4^{th}$ places respectively compared to other positions, but with decreasing frequency in that order The frequency of positions of selection of other terms was comparatively more spread out

In a similar study of quadratic predictors reported recently [139], the basis of comparison with LP is the time delay d upto which signal correlations are considered in the model rather than the number of coefficients p as above. For a time delay, $d = 10$ samples, the number of coefficients in the linear predictor is 10 while that in the quadratic predictor (including 10 linear terms) is 65 For this case, using an analysis frame length $N_f = 200$ samples (25 ms at 8 kHz sampling rate) gives a segmental prediction gain of 18 2 dB with the quadratic predictor compared to 14 1 dB of the linear predictor This corresponds to a significant improvement of 4 1 dB with a quadratic predictor model Another important observation is that the short term quadratic predictor is capable of modelling the pitch period redundancy to a great extent which is not possible with a 10^{th} order LP

The above observations suggest that while a quadratic filter may not offer significant advantage over a linear predictor in the usual forward adaptive mode in which it is used in most medium to low bit rate coders, it may give some advantages over LP when used in a backward adaptive mode (for example as in low delay, medium bit rate coders) which does not require the transmission of the model coefficients The possible advantages may be in terms of prediction gain and the better ability to model the pitch period redundancies with a short term predictor only These speculations must be seen in perspective of the 16 kb/s LD-CELP for example,

which uses a relatively large 50^{th} order *linear* predictor (in backward adaptive mode) in order to model the long term correlations in the signal as well

Chapter 5

Local State Prediction Coding of Speech

By casting a scalar time series prediction problem in state space by reconstructing a vector time series, we can use any one of two basic prediction schemes, namely, the global and local prediction schemes The reconstruction of a vector time series from a scalar observable using the method of time delays has sanction in dynamical systems theory through Takens' theorems which we discussed in Chapter 2 There exists a deep connection in the form of differentiable equivalence between the reconstructed vector time series and the original dynamical system trajectory whose time evolution is monitored through the scalar observable (whose prediction we are interested in over here) In chapter 4, we investigated a global prediction scheme for speech using polynomial representation form It is easy to see that the usual linear prediction schemes (forward adaptive, backward block adaptive and recursive prediction) can be formulated as problems of one step global prediction of linear dynamical systems in appropriately reconstructed state space (sections 1 4 and 4 3 1) In contrast to a global prediction scheme where the system parameters are optimized over the reconstructed vectors over the *entire* state space, a local prediction scheme optimizes the parameters over a *local* state space volume where the prediction is to be done Thus, a local state prediction scheme reduces the dependence on the representation form compared to global state prediction

A Local State Prediction (LSP) scheme can be implemented as follows

- Embed the scalar time series in a trajectory in state space using the method of time delays (see eg chapter 2) Call this the "reconstructed trajectory" and the dimension of the state space as the "embedding dimension", d
- For a one step prediction of a *target vector* on the reconstructed trajectory,
  - ⋆ Find a local neighbourhood of N_L trajectory points of the target vector from (possibly fixed number N_f of) previous points on the trajectory,
  - ⋆ Fit a local model between the trajectory points in the neighbourhood and their respective future points,
  - ⋆ Use this prediction model on the target vector to obtain its prediction
- Project the predicted vector on an appropriate coordinate axis depending on the method of reconstruction, to obtain the predicted scalar value

We will study the prediction properties of LSP for speech in terms of the segmental prediction gain, plots of the residual sequence, their spectrum and autocorrelation function and compare them with those of (i) short term Linear Prediction (LP) residual, and, (ii) short term plus long term LP residual In a local state prediction scheme for speech, an appropriately chosen neighbourhood will contain trajectory points that are close to the target vector in time as well as those which are approximately an integral number of pitch or formant periods away. Thus, a LSP attempts to simulate the function of both short term and long term linear predictions simultaneously

The motivation for studying the prediction properties of a LSP scheme is to explore whether it can be advantageously used in a speech coding scheme A natural method of incorporating LSP in a speech coder is to use it analogous to a backward adaptive scheme This is because the usual forward adaptive method of obtaining the optimal parameters within a frame of speech and transmitting them may prove to be extremely expensive in terms of bit rate for a local prediction scheme where optimal model fits have to be done in local neighbourhoods *within* the frame Thus, a speech coding scheme based on LSP can provide low coding delay in which context backward adaptive and recursive LP schemes are usually studied and used We propose a framework for low to medium delay speech coding in the medium bit

rate range based on LSP. The coder uses an analysis – by – synthesis scheme and is structurally similar to CELP. It is tentatively named as a Vector Excited Local State Prediction (VELSP) coder.

The organization of the chapter is as follows. In section 5.1, we discuss the steps involved in local state prediction analysis. Section 5.2 is concerned with the performance of iterative prediction of speech based on an autonomous LSP system. This prediction performance is interpreted in terms of the metric entropy results of chapter 3. We discuss in some detail, the prediction properties of a one step local state predictor for speech in section 5.3. Section 5.4 recapitulates some of the recent studies in local methods for speech prediction and coding. Finally, in section 5.5, we propose the VELSP coding scheme and discuss the performance of a skeletal structure of such a coder.

5.1 Local State Prediction (LSP) Analysis

In this section, we will elaborate on the steps given above for the implementation of LSP. The basic idea is illustrated in fig. 5.1. Let us consider that a scalar time series x_n, $n = -N_f - d + 1, \quad , -1$ is available based on which we have to predict the time series for N_p steps from $n = 0$ to $n = N_p - 1$ using LSP. Here, N_f is the *analysis frame* length and d is the *embedding dimension* of the state space in which local state prediction is to be done. The choice of embedding dimension d comes from a knowledge of the dimension of the time series and the attendant necessary and sufficient conditions for state space reconstruction (chapter 3) or from extensive experimentation with several realizations of the underlying process of the time series observable.

The first step is to reconstruct a vector time series $\mathbf{x}_n^d$, $n = -N_f, \quad , -1$, where

$$\mathbf{x}_n^d = [x_{n-d+1}\ x_{n-d+2} \quad x_{n-1}\ x_n]^T \tag{5.1}$$

The local state predictor is given by

$$\hat{\mathbf{x}}_n^d = \mathbf{g}(\hat{\mathbf{x}}_{n-1}^d) \tag{5.2a}$$

$$\hat{x}_n = \mathbf{h}^T \hat{\mathbf{x}}_n^d, \quad n = 0, 1, \quad , N_p - 1 \tag{5.2b}$$

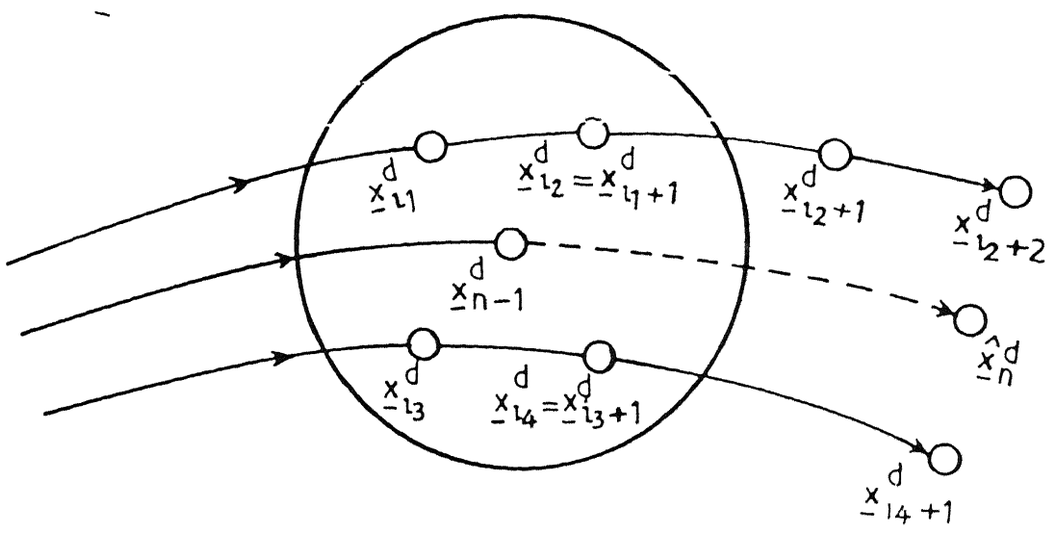


Fig. 5.1: Schematic showing local state prediction A local neighbourhood of the current trajectory point $\mathbf{x}^d_{n-1}$ is found from previous trajectory points to predict the future point $\hat{\mathbf{x}}^d_n$

Also,

$$\hat{\mathbf{x}}^d_{-1} = [\hat{x}_{-d}\ \hat{x}_{1\ d} \quad \hat{x}_{\ 2}\ \hat{x}_{\ 1}]^T$$

$$= \mathbf{x}^d_{-1} \tag{5.3a}$$

$$\mathbf{h} = [0 \quad 0\ 1]^T \tag{5.3b}$$

$$\mathbf{g}(\hat{\mathbf{x}}^d_{n-1}) = \begin{bmatrix} 0 & 1 & 0 & & 0 \\ 0 & 0 & 1 & & 0 \\ & & & & \\ 0 & 0 & 0 & & 1 \\ 0 & 0 & 0 & & 0 \end{bmatrix} \hat{\mathbf{x}}^d_{n\ 1} + \begin{bmatrix} 0 \\ 0 \\ \\ 0 \\ 1 \end{bmatrix} f(\hat{\mathbf{x}}^d_{n\ 1})$$

$$= \left[\hat{x}_{n-d+1}\ \hat{x}_{n\ d+2} \quad \hat{x}_{n\ 1}\ f(\hat{\mathbf{x}}^d_{n\ 1})\right]^T \tag{5.3c}$$

Here, $\mathbf{g}(\hat{\mathbf{x}}^d_{n-1})$ is the local state predictor of $\hat{\mathbf{x}}^d_{n-1}$ and $f(\hat{\mathbf{x}}^d_{n\ 1})$ is a local representation form

In order to evaluate $f(\hat{\mathbf{x}}^d_{n-1})$, we need to define a N_L point *neighbourhood* of $\hat{\mathbf{x}}^d_{n\ 1}$ A simple way to assign neighbourhoods is to partition the state space into disjoint

sets using, for example, a rectangular grid. While such an approach is convenient and efficient in the case of successive predictions using the same analysis frame, it has the disadvantage that there is no overlap between neighbourhoods. A point near the boundary of its neighbourhood may be poorly approximated One way to cope with this problem is to perform interpolation between optimal functions of adjacent neighbourhoods However, this becomes a difficult task for state space dimension greater than two An alternative that is more accurate than disjoint partitions and more convenient than enforcing matching conditions at the boundaries is to overlap the boundaries in such a way that each prediction is done from a good set of neighbours

For predicting a point $\mathbf{x}_n^d$, we choose a N_L point *nearest* neighbourhood in the sense of minimum Euclidean distance from the available reconstructed trajectory points, $\mathbf{x}_n^d$, $n = -N_f, \quad , -2$ Let these N_L points and their future points be denoted pairwise as $(\mathbf{x}_{i_1}^d, \mathbf{x}_{i_1+1}^d)$, $(\mathbf{x}_{i_2}^d, \mathbf{x}_{i_2+1}^d)$, $\quad (\mathbf{x}_{i_{N_L}}^d, \mathbf{x}_{i_{N_L}+1}^d)$

The next step is to obtain the parameters of a local representation form which maps the N_L points to their respective future points in some optimal sense Various representation forms can be chosen for this purpose However, the dependence of the prediction performance on representation can be expected to be less for LSP compared to global prediction A trivial choice of a local representation is a *constant* map The next higher level of approximation, which we have studied for speech, is provided by a local linear predictor Such a predictor for a state space point $\mathbf{x}^d$ is given by

$$f(\mathbf{x}^d) = a_0 + [a_1 \quad a_d]\,\mathbf{x}^d \tag{5 4}$$

where $a_0, \quad , a_d$ are the $d+1$ predictor coefficients If the number of nearest neighbours $N_L = d+1$, then one can simply do linear interpolation to find the coefficients However, one is generally interested in the case $N_L > d$, which provides an overdetermined set of linear equations in the predictor coefficients

$$x_{i_j+1} = a_0 + [a_1 \quad a_d]\,\mathbf{x}_{i_j}, \quad j = 1, \quad , N_L \tag{5 5}$$

where x_{i_j+1} is the d^{th} scalar component of $\mathbf{x}^d_{i_j+1}$ The $(d+1)$ coefficients are then solved from a set of simultaneous linear equations which are obtained by doing a *weighted* minimization of E with respect to $a_0, a_1, \quad , a_d$, where

$$E = \frac{\sum_{j=1}^{N_L} w_{i_j}^2 \left[x_{i_j+1} - f(\mathbf{x}^d_{i_j}) \right]^2}{\sum_{j=1}^{N_L} w_{i_j}^2} \tag{5.6a}$$

and

$$w_{i_j} = \frac{1}{\|\mathbf{x}^d - \mathbf{x}^d_{i_j}\|_2}, \quad j = 1, \quad , N_L \tag{5.6b}$$

The local linear predictor (LLP) can be seen as a generalization of the usual linear predictor (LP) A LP based on d delays can be considered as a d-dimensional plane *passing through the origin* in $(d+1)$ dimensional space, where the d axes correspond to the d delay coordinates and the $(d+1)^{th}$ axis gives the prediction coordinate A corresponding LLP can be considered as a *surface* S in a $(d+1)$ dimensional plane The scalar prediction $\hat{x}_n$ of the point $\hat{\mathbf{x}}^d_{n-1}$ will be given by the $(d+1)^{th}$ coordinate of the corresponding $(d+1)$ dimensional point on S The LLP $f(\mathbf{x}^d_{n-1})$, eq (5.4), will be given by a plane tangent to the surface at that point The additional constant, a_0, in the predictor accounts for any local mean of the reconstructed trajectory, $\mathbf{x}^d_n$ Also note that if the local neighbourhood extends to include the entire state space, then a LLP reduces to the usual linear predictor

In the following two sections, we study the properties of many step, i e , iterative, and one step prediction of speech using local linear state prediction

5.2 Performance of an Autonomous Local State Prediction System for Speech

We will first look into the prediction performance of an *autonomous* local state prediction system for speech, i e, one which does not get external excitation after starting the iteration from an initial condition To begin with, let us consider a speech frame, x_n, $n = -N_f - d + 1, \quad , -1$, is available The first step is to reconstruct a vector trajectory $\mathbf{x}^d_n$, $n = -N_f, \quad , -1$ in d- dimensional state space using eq (5.1) We will refer to this reconstructed trajectory of length N_f as the *analysis frame*

The next step is to predict N_p samples $\hat{x}_n$, $n = 0, \quad , N_p - 1$ using eqs (5 2)–(5 6) If the corresponding speech samples are given by x_n, $n = 0, \quad , N_p - 1$, then

$$e_{n,\iota} = x_{n,\iota} - \hat{x}_{n,\iota}, \quad n = 0, 1, \quad , N_p - 1 \tag{5 7}$$

denotes the prediction error at iteration step $n+1$ The additional subscript ι is to denote the analysis frame number Successive N_p step predictions are performed by shifting the analysis frame by N_p points on the speech time series The above prediction scheme is shown schematically in fig 5 2. It is expected that the average prediction error is a function of the iteration step — as one predicts further in time, the prediction based on the analysis frame and the previous reproduced points is expected to get worse We quantify this, for the case of speech, by plotting the prediction gain, $PG(n)$, in dB, as a function of the iteration step n, in fig 5 3 Here,

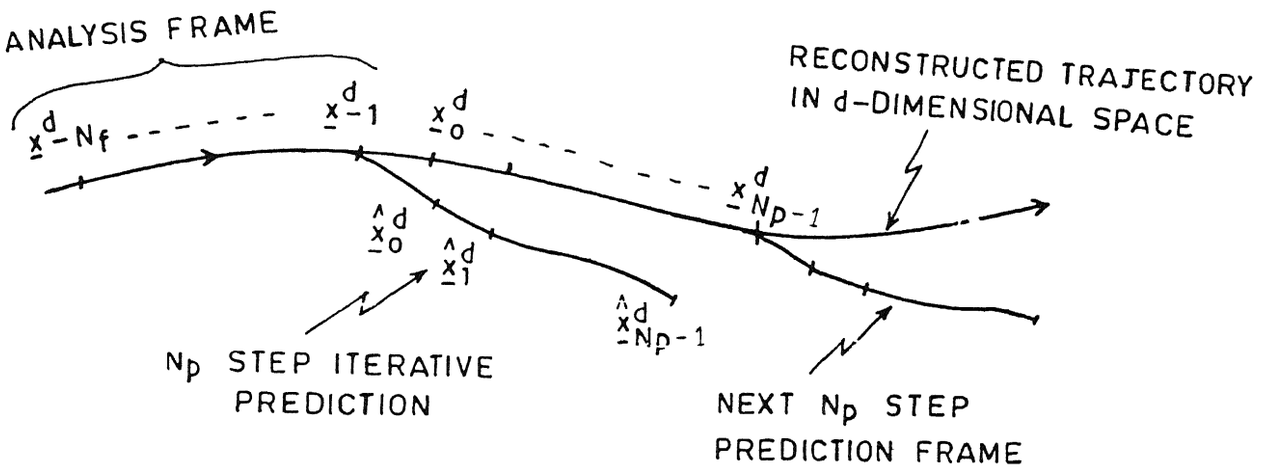


Fig. 5.2: An N_p step iterative prediction using a local state predictor

$$PG(n) = 10\log_{10}\frac{\sum_i e_{n,i}^2}{\sum_i x_{n,i}^2}, \quad n = 0, 1, \quad , N_p - 1 \tag{5 8}$$

We have used the 4 sentence utterances of speech database 2 (Appendix B) by 3 male and 3 female speakers to obtain the plot of fig 5 3 The other parameter choices are analysis frame length $N_f = 160$, number of nearest neighbours $N_L = 40$, embedding dimension $d = 10$ and number of iteration steps $N_p = 10$ An *autonomous* local state prediction system is unsuitable for use as such in a speech coder because of the rapid decrease in the prediction gain with increasing iteration steps Even if we allow an excitation function for the prediction system, in order to get a prediction gain independent of the iteration step number, one has to incorporate an asymmetric dependence of the excitation function on the iteration step which is undesirable

The graph of fig 5 3, however, serves to verify the results of second order dynamical entropy, K_2, for speech reported in chapter 3 Recall that for an *autonomous* dynamical system, the metric entropy K quantifies the rate of loss of information about its initial condition Also, $K \geq K_2$ If R is the resolution of the initial condition in bits, then the dynamical system can be iterated for $N_T \sim \frac{R}{K}$ steps on the average, before all information about the initial condition is lost and no further iteration is possible For $K_2 = 0\,54$ bits/sample (8613 s^{-1} at 16 kHz – see Table 3 2), and $R = 16$ bits, $N_T < 29\,7$ iterations It is seen from fig 5 3 that the average prediction gain with an autonomous LSP system goes below 0 dB after 9 iterations on the average The difference in the two results can be largely attributed to the fact that while the former is a theoretical estimate, the latter is a model specific result Nonetheless, the comparative values reinforce our faith in the second order entropy results for speech time series

5.3 One Step Local State Prediction of Speech

In this section, we study in some detail the one step local state predictor for speech For the one step case, $N_p = 1$ in the LSP analysis formulation of section 5 1 (eq (5 2)) Based on length $N_f + d - 1$ of scalar data, a d-dimensional vector time series of length N_f is reconstructed Next, a one step prediction into the future is done using

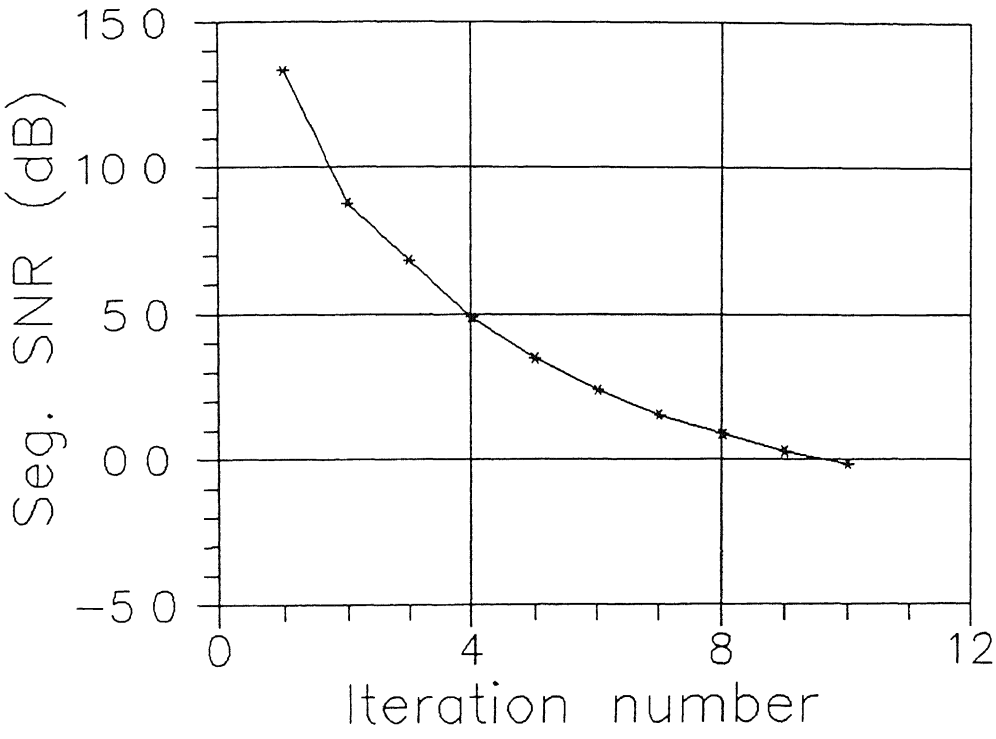


Fig. 53. Prediction gain vs iteration step using an autonomous local state predictor system

LSP The subsequent analysis frame is obtained by shifting the previous frame by one sample and another one step prediction is performed This process is repeated for the entire time series

Given a scalar time series, x_n, $n = -N_f - d + 1, \quad , -1, 0, 1, \quad , N-1$, the first $N_f + d - 1$ samples are used to form the first analysis frame based on which $\hat{x}_0$ is predicted Let the predicted sequence be $\hat{x}_n$, $n = 0, 1, \quad , N-1$ The predicted point $\hat{x}_i$, $i = 0, 1, \quad , N-1$ uses an analysis frame of scalar points x_n, $n = i - (N_f + d - 1), \quad , i-1$ There are various factors in a local state prediction scheme which can affect its prediction performance These are the *vector* analysis frame length N_f, the embedding dimension d and the number of nearest neighbours N_L, apart from the representation form for the local fit For a local linear representation, we have found the segmental prediction gains (henceforth referred to as segmental SNR) for each combination of N_f, d and N_L from $N_f = 160, 240, 320$ $d = 1, 4, 7, 10$ and $N_L = 10, 20, 30, 40, 50$ The results are summarized in figs 5 4(a)–(e) and fig 5 5 The segmental SNR reported in this chapter (and elsewhere) are based on lengths of 160 samples unless explicitly stated otherwise The speech database used for this study consists of the 4 phoneme – specific sentences

(a) Why were you away a year, Roy?

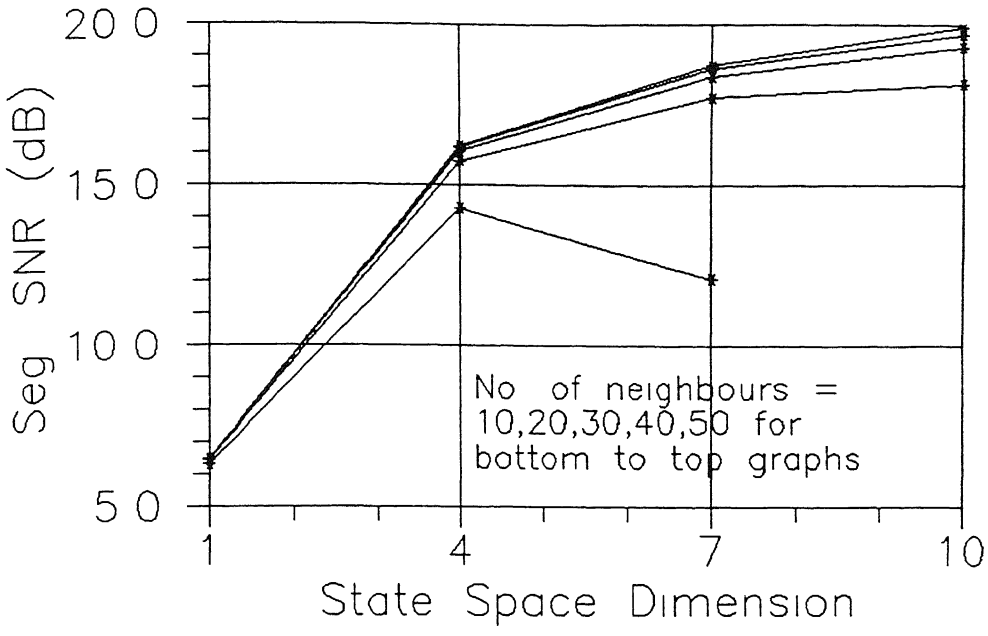
(b) Nanny may know my meaning

(c) His vicious father has seizures

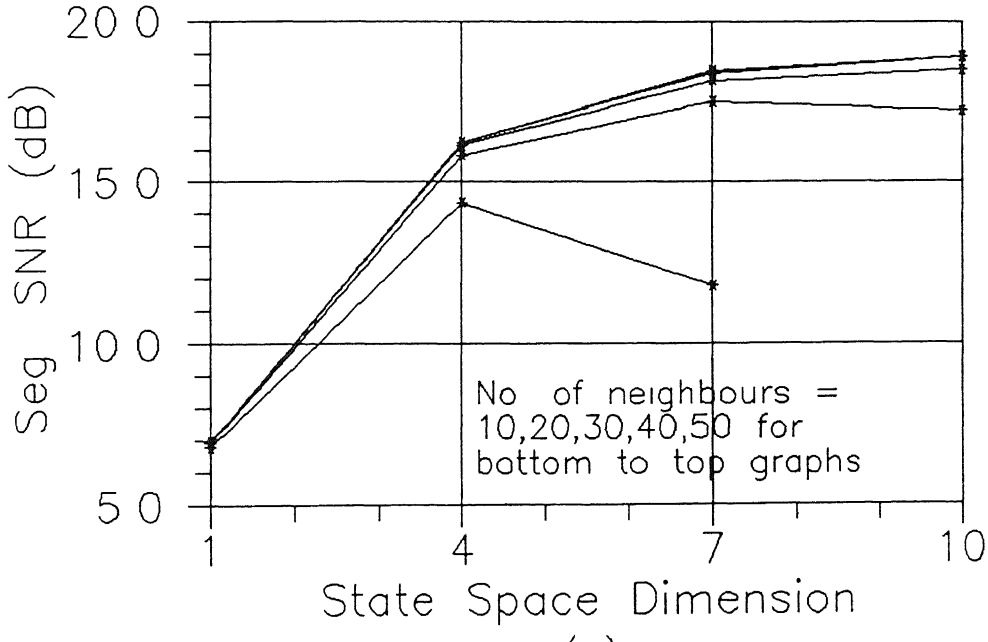
(d) Which tea party did Baker go to?

Each sentence was spoken by 3 males and 3 females (speech database 2, Appendix B) Figures 5 4(a)–(d) show for the 4 sentences (a)–(d) respectively, the segmental SNR in dB as a function of d for $N_L = 10, 20, 30, 40$ and 50 each and $N_f = 160$ Figure 5 4(e) shows a plot of the above averaged over all the 4 sentences The variation of the segmental SNR with the analysis frame length N_f is shown for $d = 10$ and $N_L = 20, 30, 40$ and 50 in fig 5 5

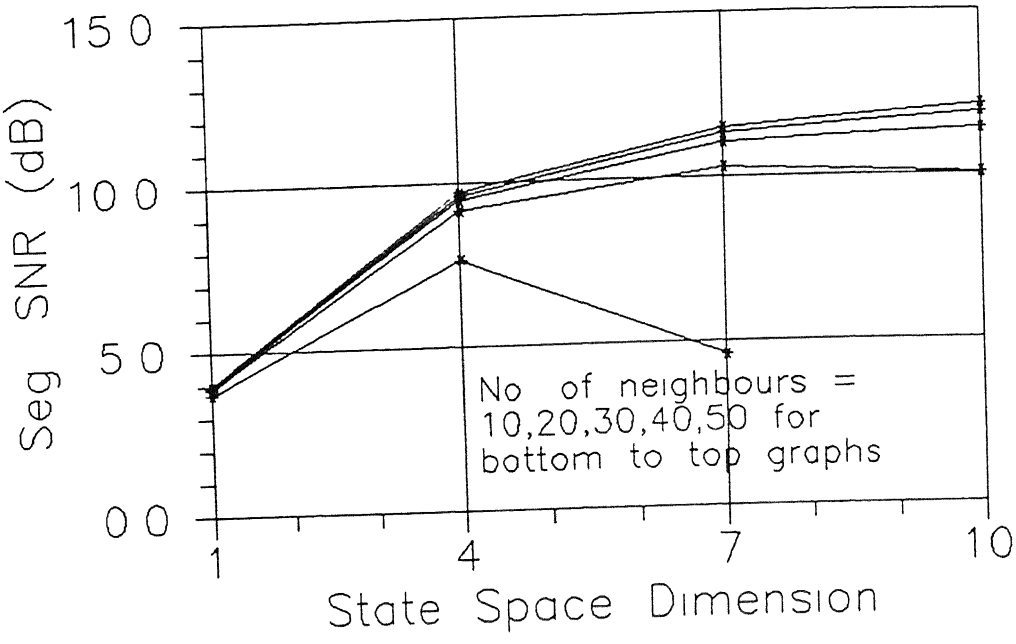
The following observations regarding the segmental SNR can be made


20 0
15 0
10 0
5 0
Seg SNR (dB)
No of neighbours =
10,20,30,40,50 for
bottom to top graphs
1
4
7
10
State Space Dimension

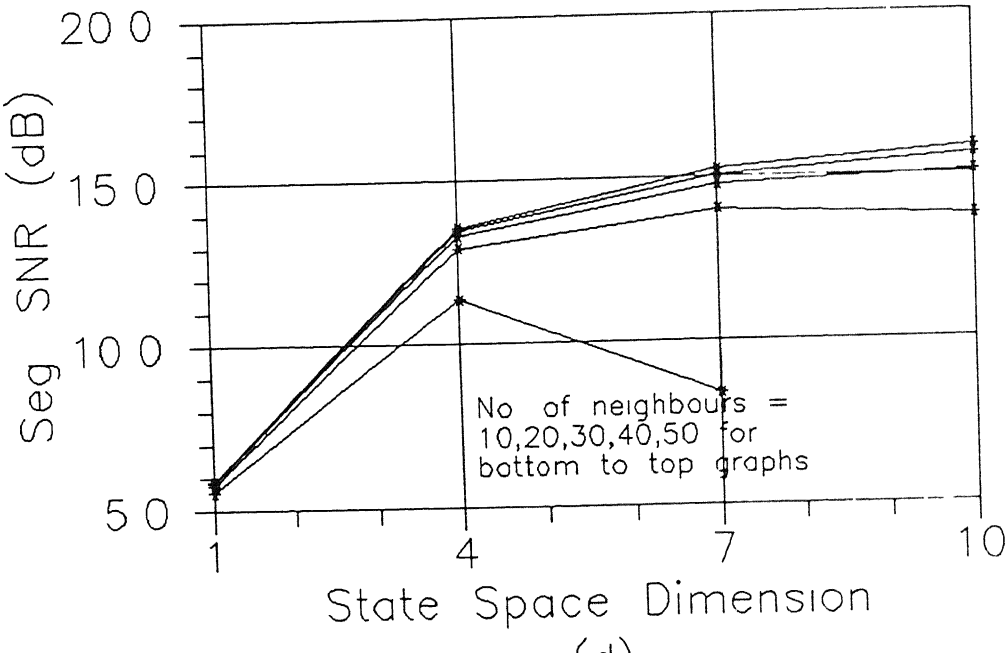
(a)


20 0
15 0
10 0
5 0
Seg SNR (dB)
No of neighbours =
10,20,30,40,50 for
bottom to top graphs
1
4
7
10
State Space Dimension

(b)


15 0
10 0
5 0
0 0
Seg SNR (dB)
No of neighbours =
10,20,30,40,50 for
bottom to top graphs
1
4
7
10
State Space Dimension

(c)


20 0
15 0
10 0
5 0
Seg SNR (dB)
No of neighbours =
10,20,30,40,50 for
bottom to top graphs
1
4
7
10
State Space Dimension

(d)

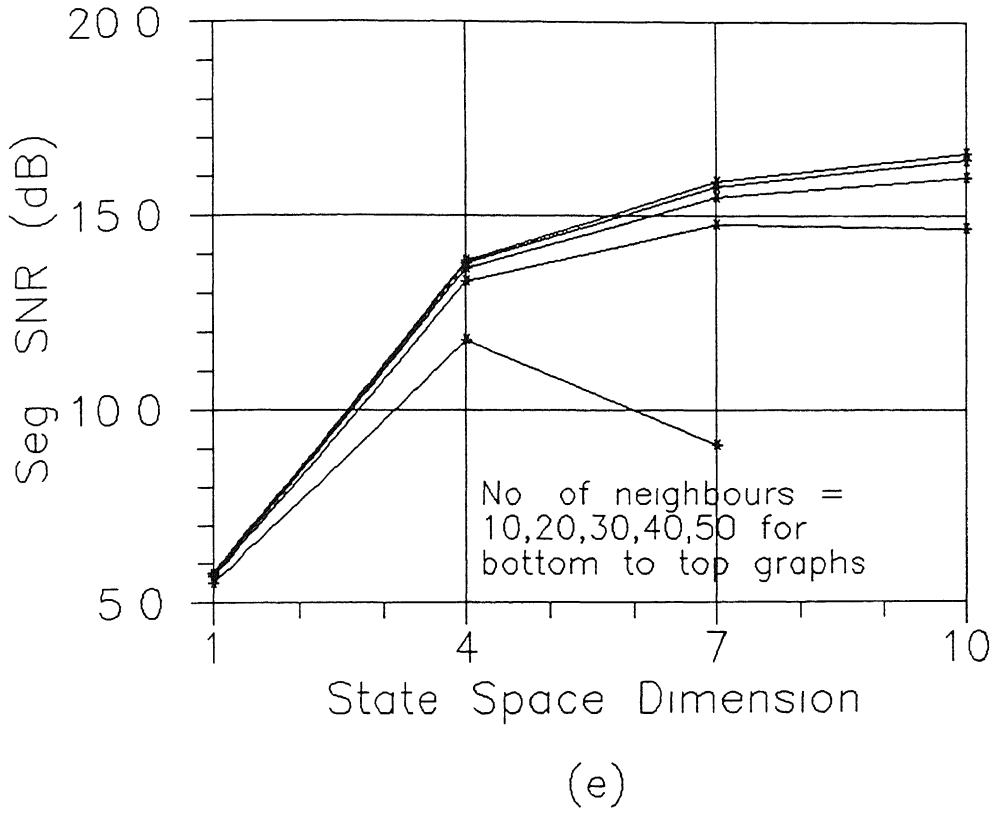


Fig 5.4: Graphs showing segmental prediction gain in dB vs the state space or embedding dimension d for a local state predictor Parts (a)–(d) correspond to the 4 sentence utterances (see text) and part (e) gives the overall average Each graph shows 5 plots corresponding to local neighbourhood size, N_L = 10, 20, 30, 40 and 50 The analysis frame length N_f = 160 samples

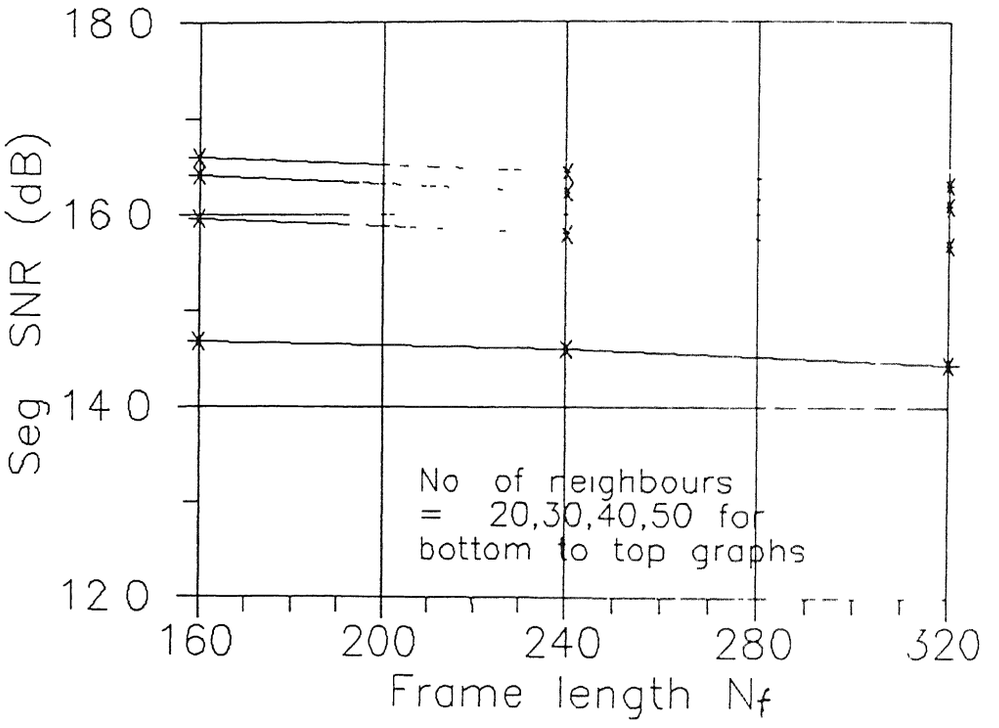


Fig. 5.5: Segmental prediction gain in dB as a function of the analysis frame length N_f for a local state predictor The 4 plots correspond to local neighbourhood size N_L= 20, 30, 40 and 50 In each case, the embedding dimension, d = 10

(1) Comparing the graphs of fig 5 4 (a)–(d), it is seen that the segmental SNR is maximum for sentence (a) which comprises of vowels and glides only The segmental SNR is minimum for the utterance of sentence (c) which contains only voiced and unvoiced fricatives besides vowels.

(2) One can compare the segmental SNR obtained for the 4 sentence utterances and for the overall database using LSP in figs. 5 4 (a)–(e) with that obtained using short term LP (covariance method) in figs 4.6 (a)–(e) It must be noted that the LP scheme optimizes the predictor parameters *within* an analysis frame (of 160 samples in this case) and the prediction is done within the frame whereas a LSP scheme predicts a point based on analysis frame consisting of *past* samples While the LP is a forward block adaptive scheme, the one step LSP is reminiscent of recursive adaptation Comparing the segmental SNRs for the entire database (figs 5 4(e) and 4.6(e)), we see that LSP performs marginally better than LP for $d = 4, 7$ and 10 and $N_L = 40$ and 50 in LSP. In both cases, $N_f = 160$ samples For $d = 10$ and $N_L = 40$, for example, the segmental SNR using LSP is 16.4 dB The comparative figure for LP is 16 08 dB

(3) The segmental SNR using LSP with a local neighbourhood of $N_L = 10$ samples is quite poor at all reconstruction dimensions. The prediction performance for $N_L = 30, 40$ and 50 is relatively similar This can be seen from figs 5 4(a)–(e)

(4) The relative increase in the segmental SNR becomes less with increasing state space dimension This increase is only 0 68 dB as d is increased from 7 to 10 for $N_L = 40$ (fig. 5 4(e))

(5) It is seen from fig 5 5 that the segmental SNR decreases slightly as the analysis frame length is increased from 160 to 320 samples This decrease may be attributed to the nonstationary nature of the speech signal *This suggests that predictive schemes based on local methods must use an adaptive analysis frame of* $N_L \sim 160$ *samples rather than a fixed frame consisting of very large number of speech samples*

In order to get a better understanding of the prediction properties of a local state predictor, we have compared the LSP residual of segments of running speech

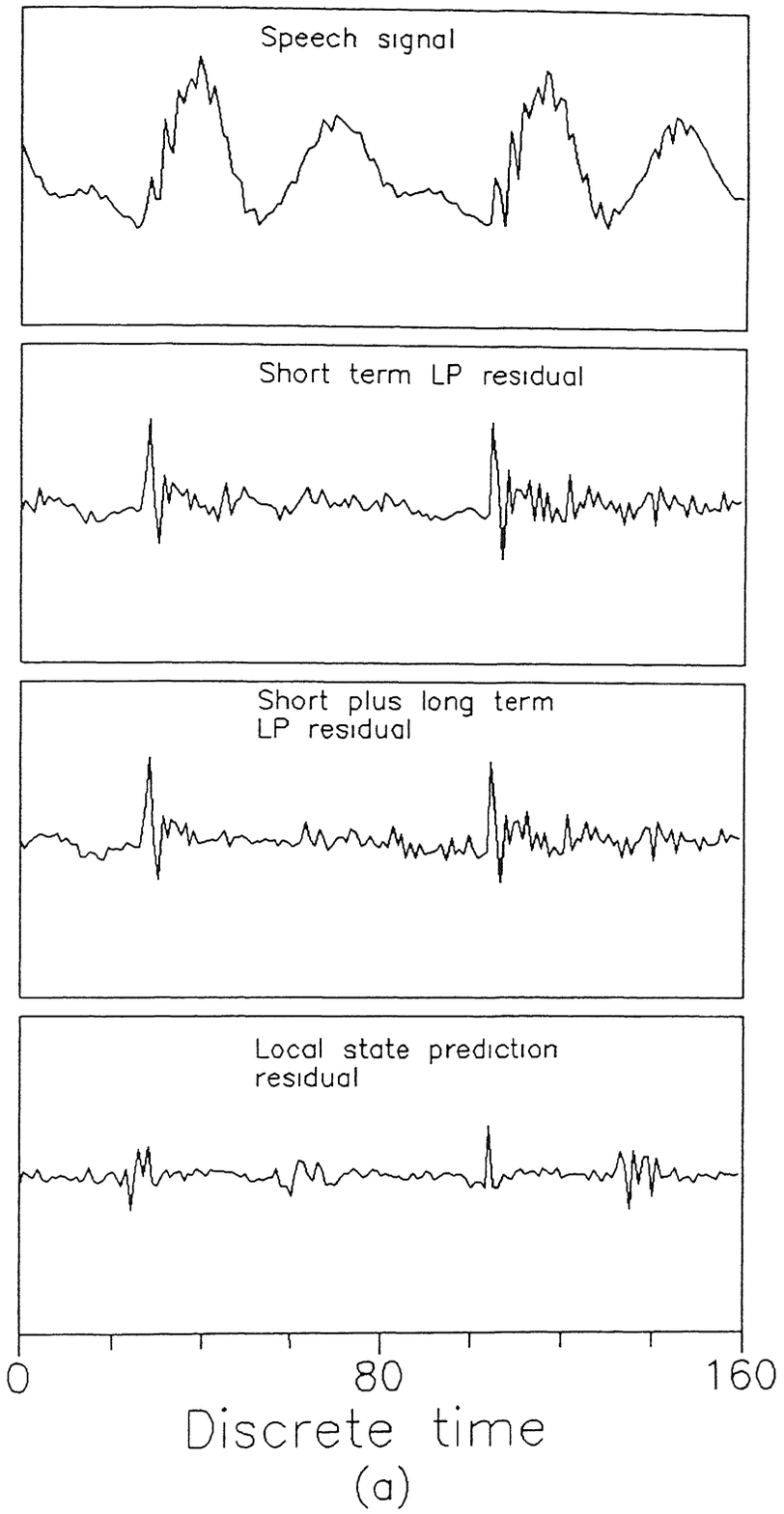
with the short term LP residual and the short term plus long term LP residual in figs 5.6 – 5.12. This comparison is done in terms of the time series plots, their autocorrelation functions and the spectrum in parts (a), (b) and (c) respectively of each figure. Based on such studies, the following general observations about the prediction scheme can be made:

(1) The performance of a LSP can be broadly said to lie between a short term LP and a short term plus long term LP both in terms of segmental prediction gain and the ability of a LSP to remove correlations in the speech signal. This can be expected because local state prediction is based on a local neighbourhood of the *target* vector which contains reconstructed trajectory vectors that are close to it in time as well as those which are close to it in space. The vectors which are close to the target vector in space but not in time are those which lie approximately an integral number of pitch or formant frequencies away from it. This gives a LSP its ability to model long term signal correlations. It is also worth noting that compared to a long term linear prediction filter which explicitly uses a pitch predictor, no hard decision regarding the pitch period has to be made in a local state predictor.

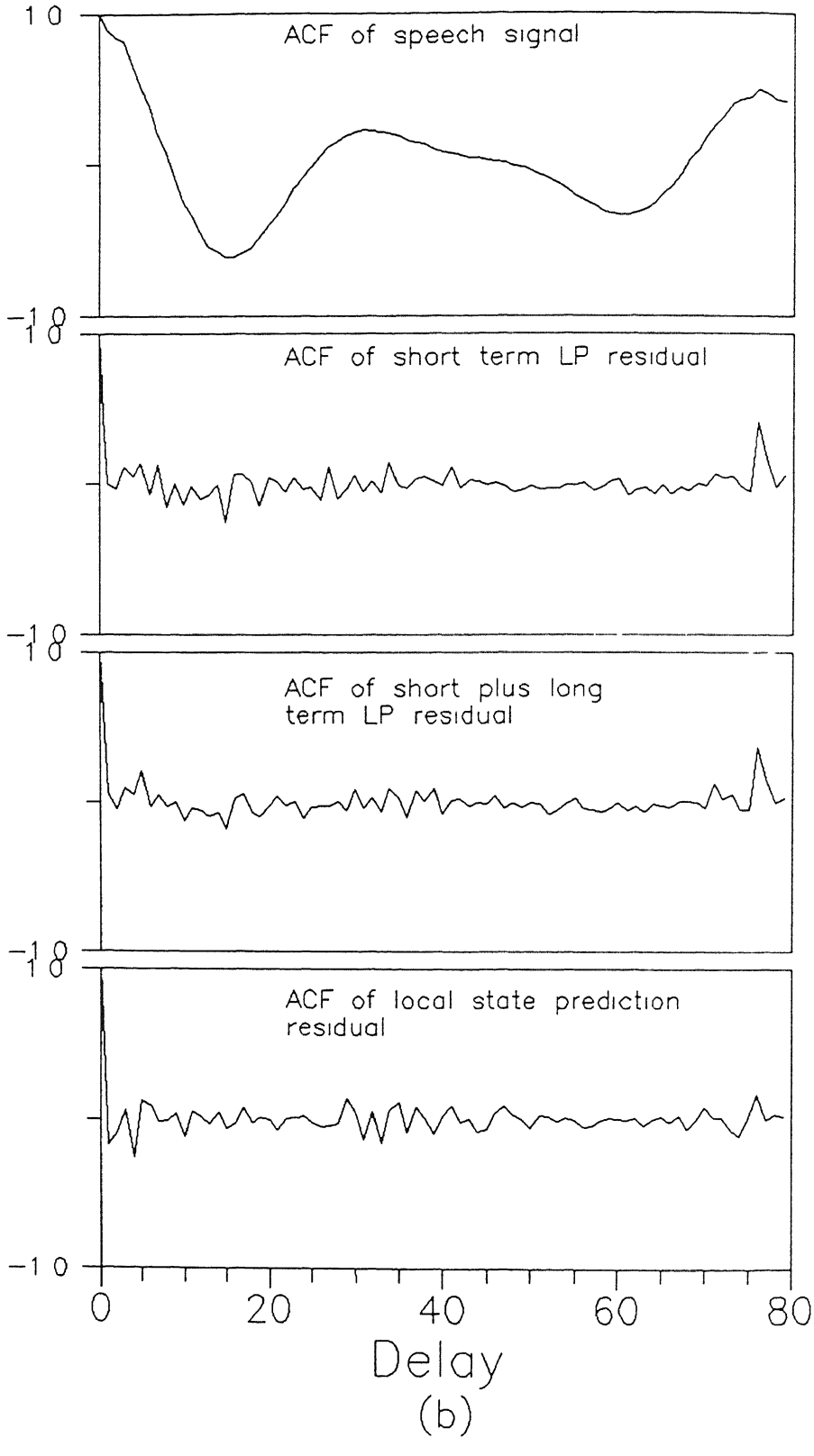
(2) The spectrum of a LSP residual appears "whiter" than that of a short term LP residual. The LSP is comparatively better at removing spectral peaks in the lower frequency range (< 1000 Hz) than at the higher frequencies.

(3) It has been observed that if the spectrum of the LSP residual contains some prominent peak, then this spectral content is generally not removed by long term linear prediction as well.

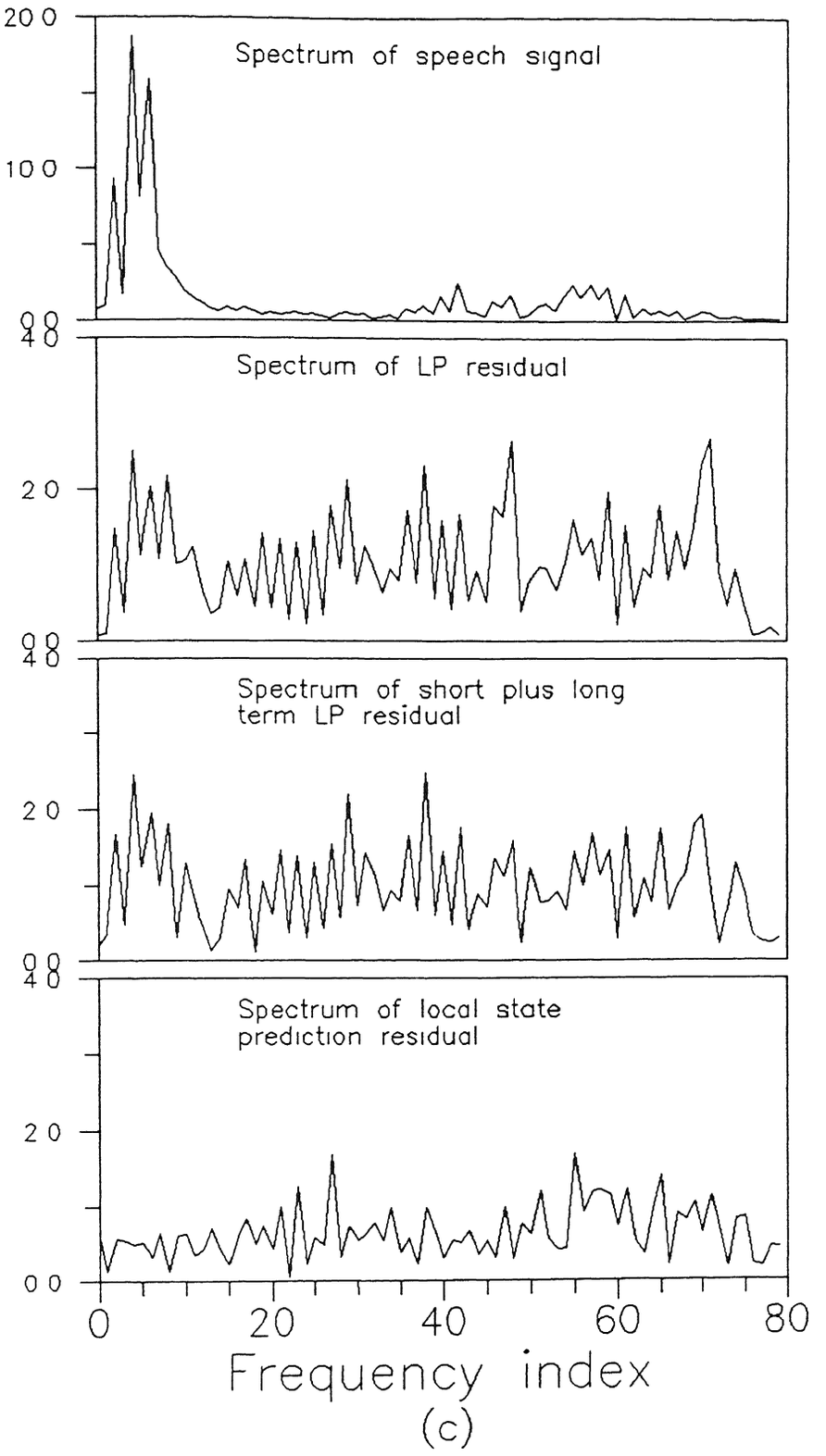
(4) Like other backward adaptive schemes, the major disadvantage of a LSP scheme lies in its inability to track sudden changes in signal characteristics compared to forward adaptive schemes. Such an example is shown in fig. 5.13 where a comparison of the LSP residual with the forward adaptive short term LP and short plus long term LP residuals is also made. In another example, Table 5.1 shows comparative SNR values of successive frames based on 160 samples each. The first frame denotes almost silent region. There is a sudden increase in the



Time series plots – 10^{th} order LP used to remove short term correlations from the speech signal The long term LP uses 3 coefficients The LSP parameters are $N_f = 320$, $N_L = 40$ and $d = 10$ The SNR based on 160 samples for the short term LP residual, short term plus long term LP residual and LSP residual are 14 58 dB, 15 23 dB and 19 16 dB respectively

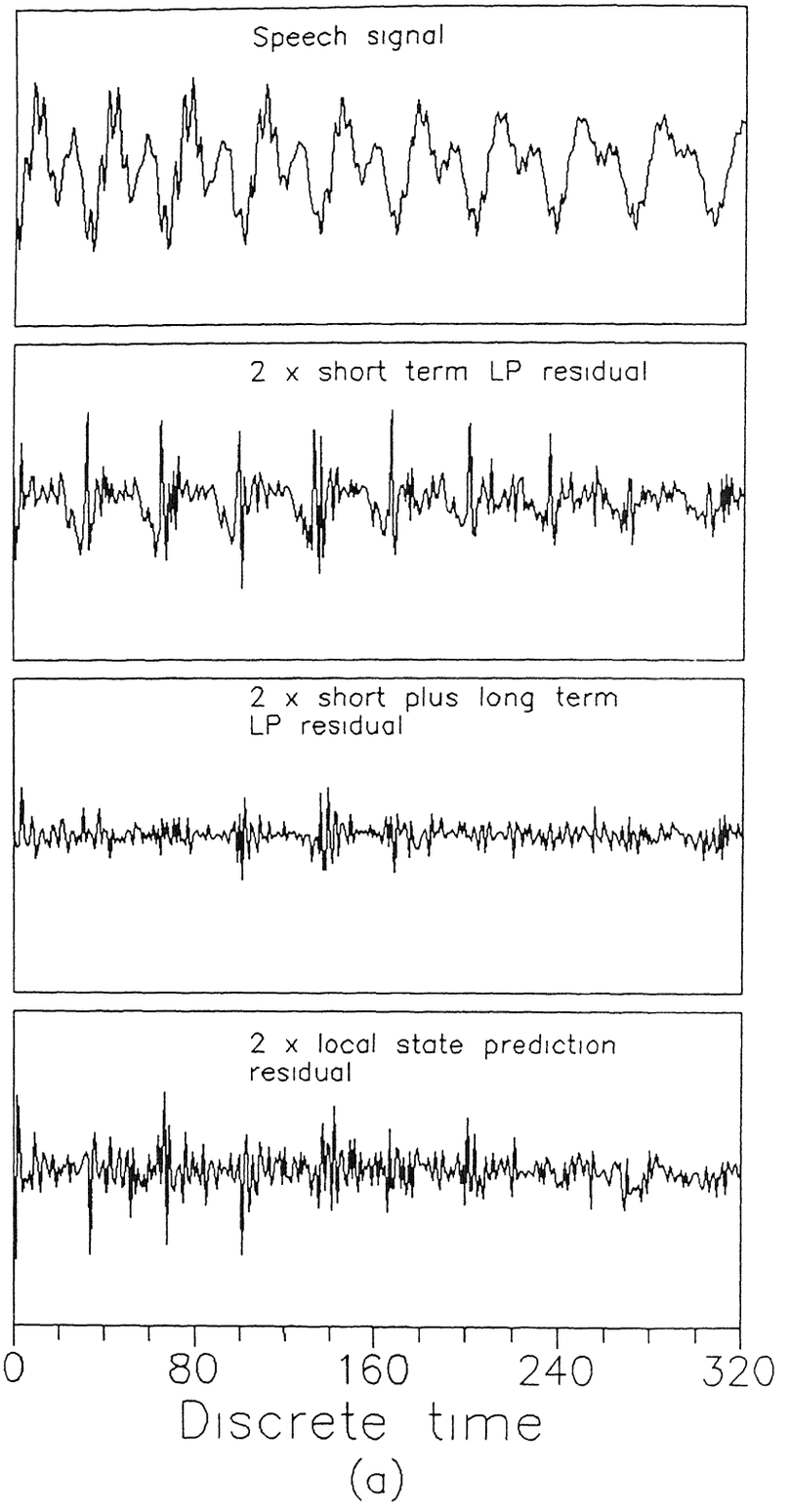


Autocorrelation function plots corresponding to the 4 time series plots of part (a)

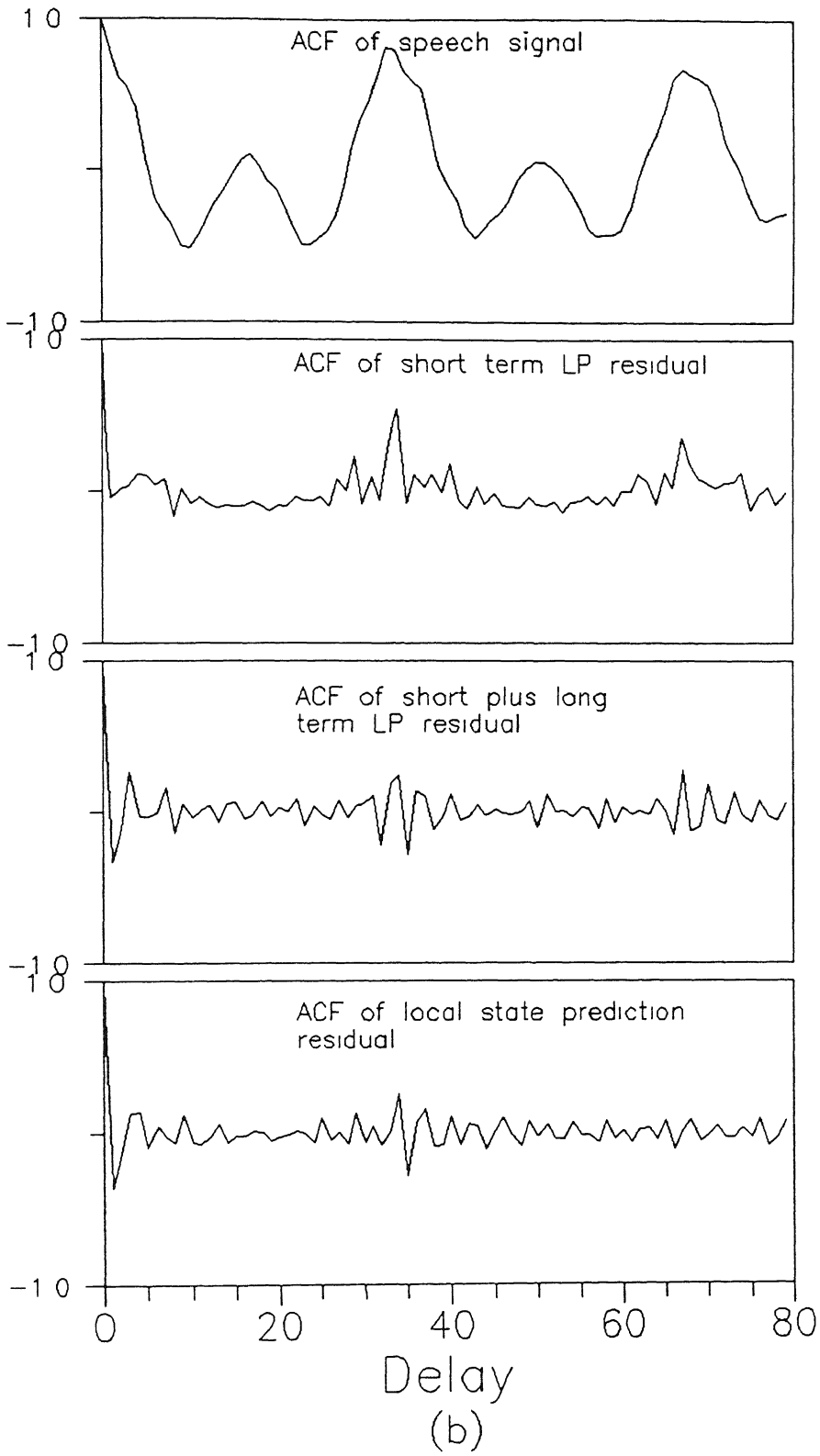


Spectrum plots corresponding to the 4 time series plots of part (a) computed using 160 point DFT Note the relative magnitudes of the plots

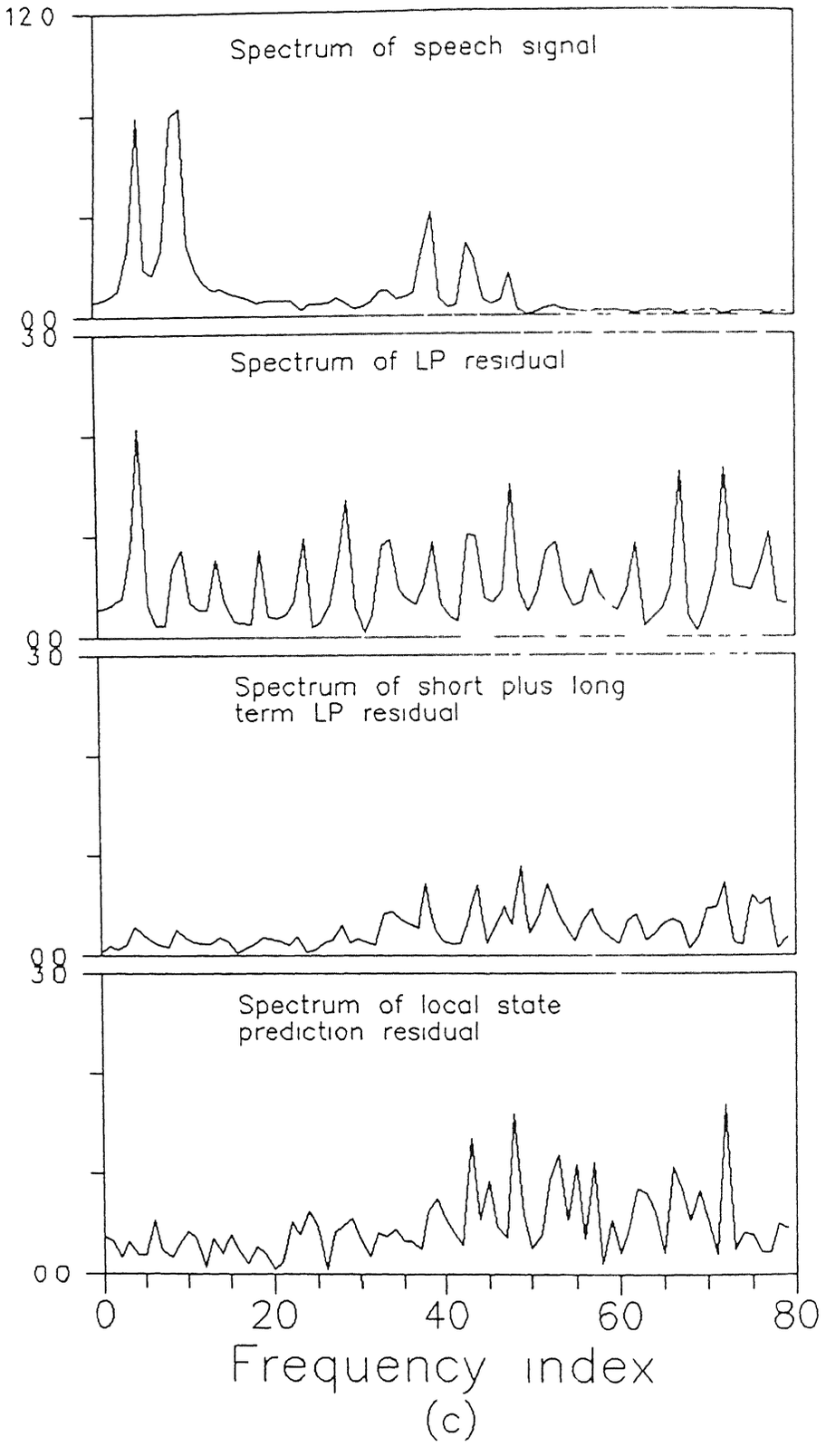
Fig. 5 6. Comparative plots of the speech signal, short term LP residual, short term plus long term LP residual and local state prediction residual in terms of (a) time series, (b) autocorrelation function, and, (c) DFT spectrum



Time series plots – 10^{th} order LP used to remove short term correlations from the speech signal The long term LP uses 3 coefficients The LSP parameters are N_f = 320, N_L = 40 and d = 10 The SNR based on 160 samples for the short term LP residual, short term plus long term LP residual and LSP residual are 15 69 dB, 21 53 dB and 17 31 dB respectively

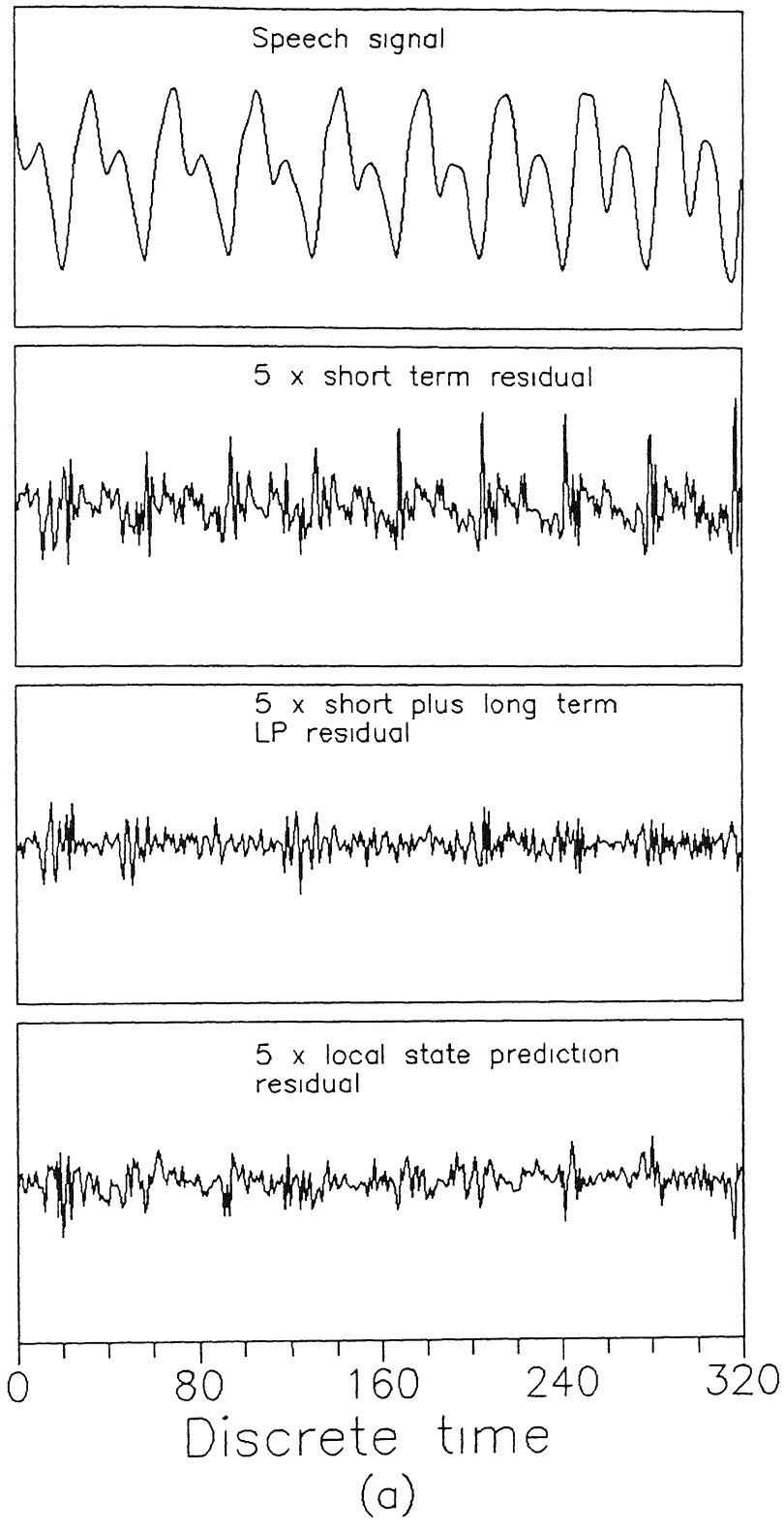


Autocorrelation function plots corresponding to the 4 time series plots of part (a)

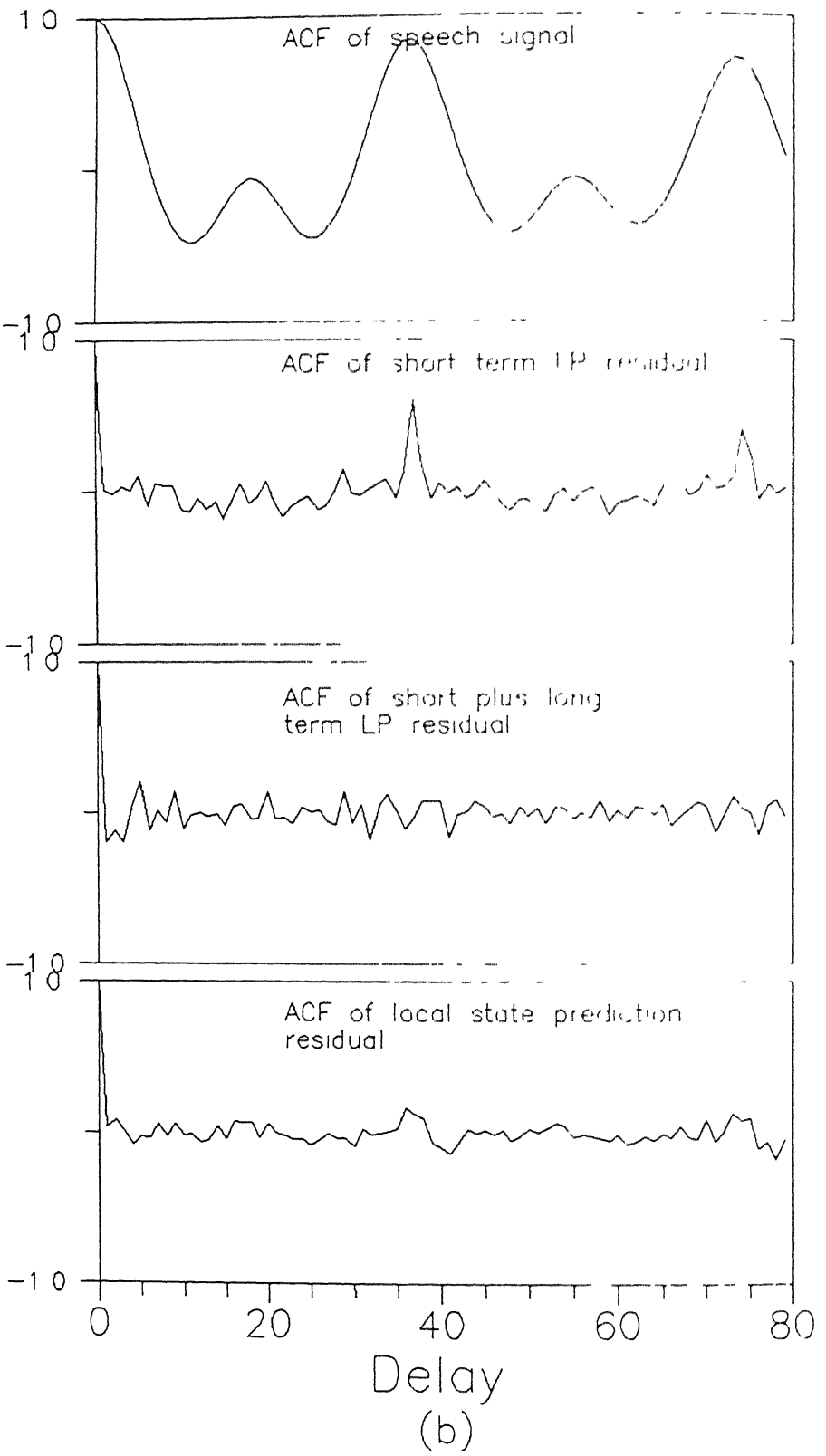


Spectrum plots corresponding to the 4 time series plots of part (a) computed using 160 point DFT Note the relative magnitudes of the plots

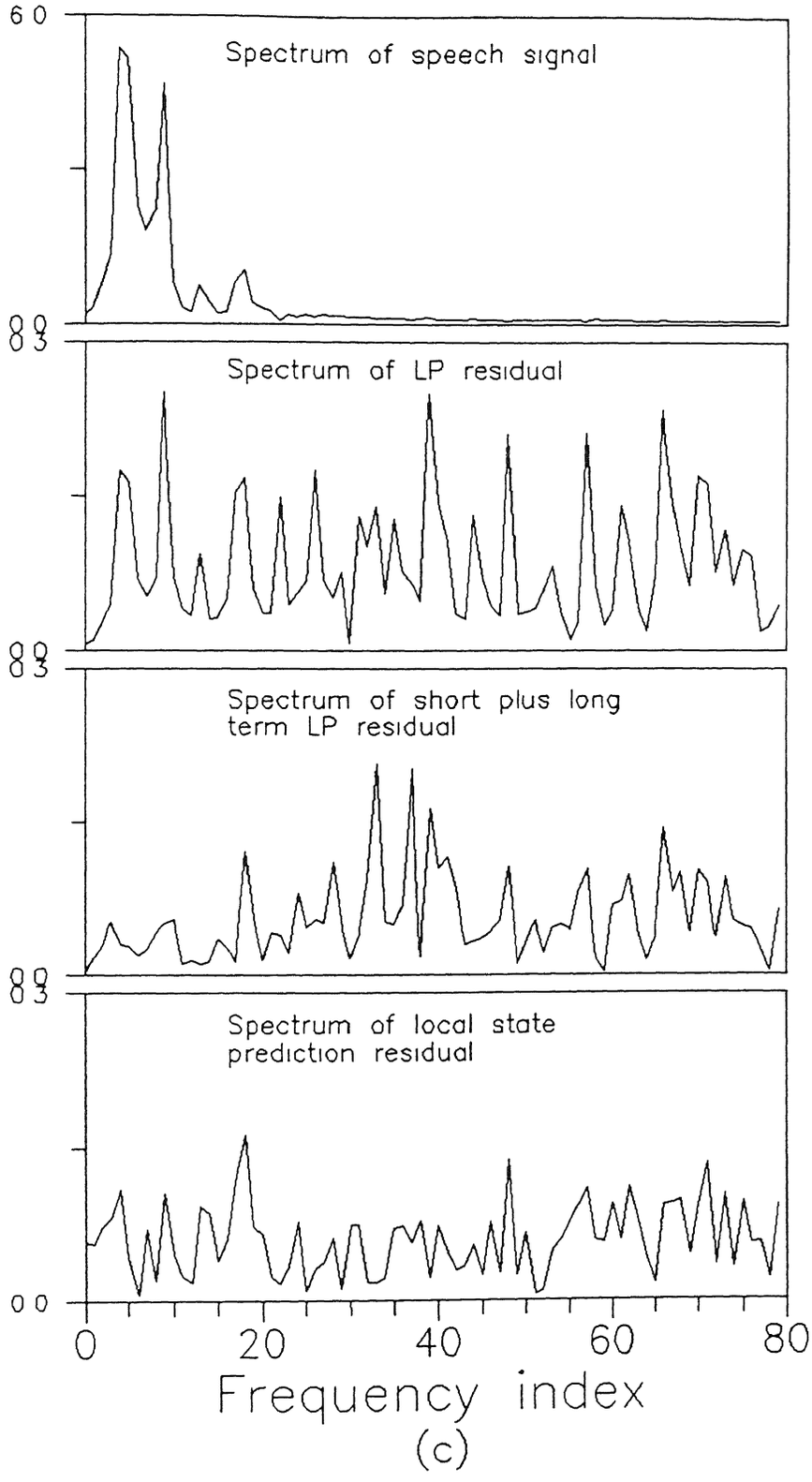
Fig. 5 7: Comparative plots of the speech signal, short term LP residual, short term plus long term LP residual and local state prediction residual in terms of (a) time series, (b) autocorrelation function, and, (c) DFT spectrum



Time series plots – 10^{th} order LP used to remove short term correlations from the speech signal The long term LP uses 3 coefficients The LSP parameters are $N_f = 320$, $N_L = 40$ and $d = 10$ The SNR based on 160 samples for the short term LP residual, short term plus long term LP residual and LSP residual are 26 55 dB, 31 48 dB and 31 10 dB respectively

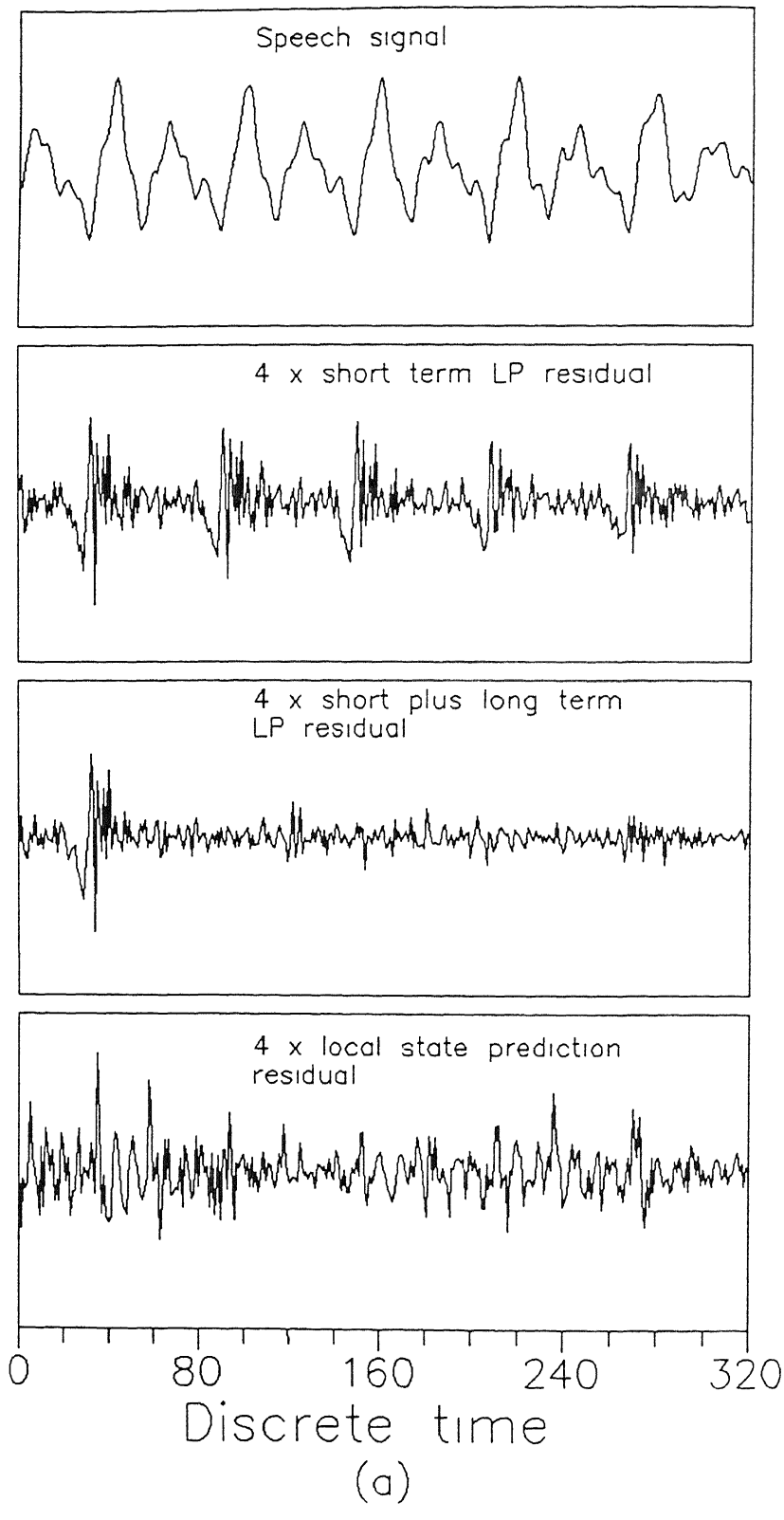


Autocorrelation function plots corresponding to the 4 time series plots of part (a)

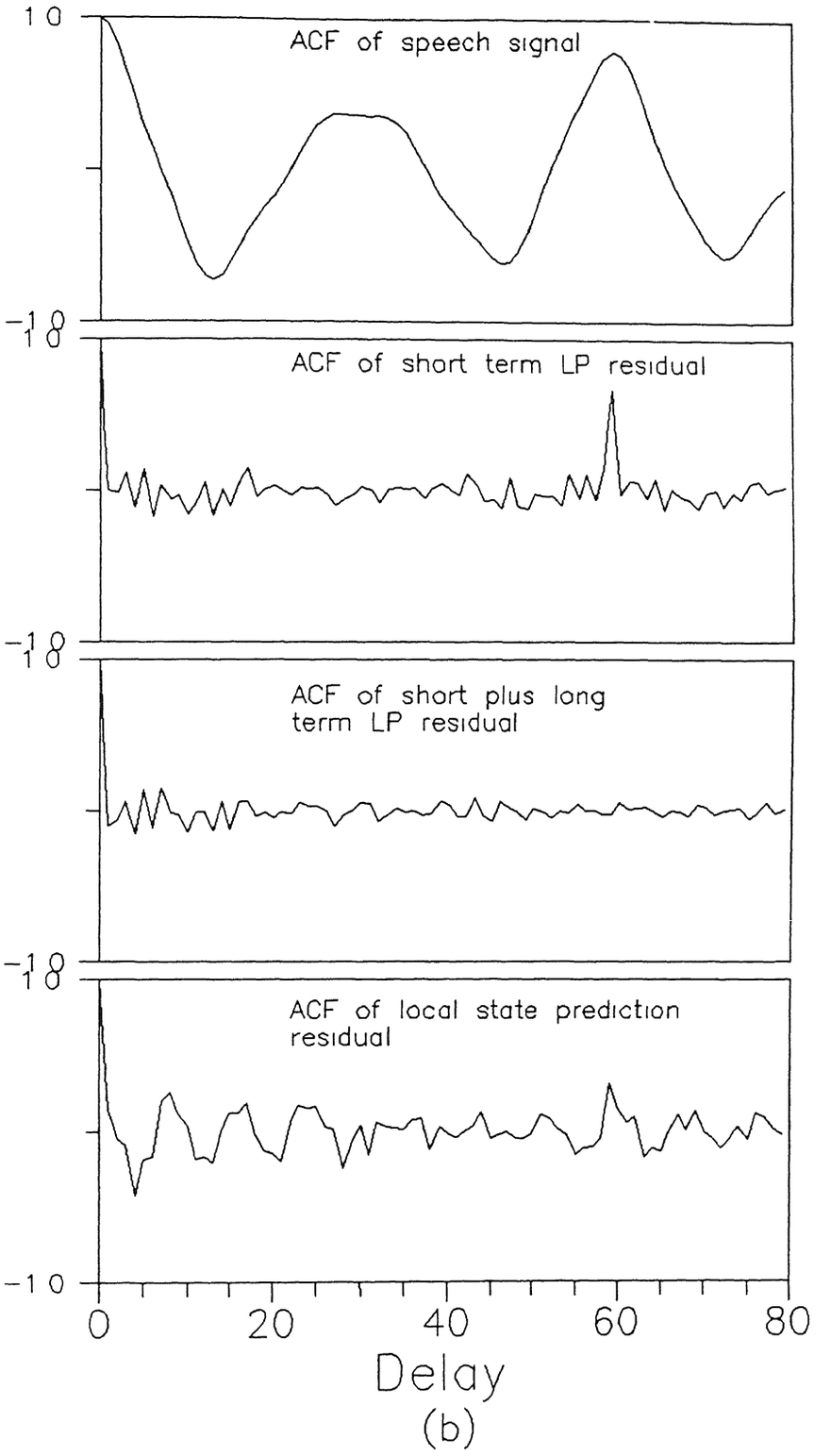


Spectrum plots corresponding to the 4 time series plots of part (a) computed using 160 point DFT Note the relative magnitudes of the plots

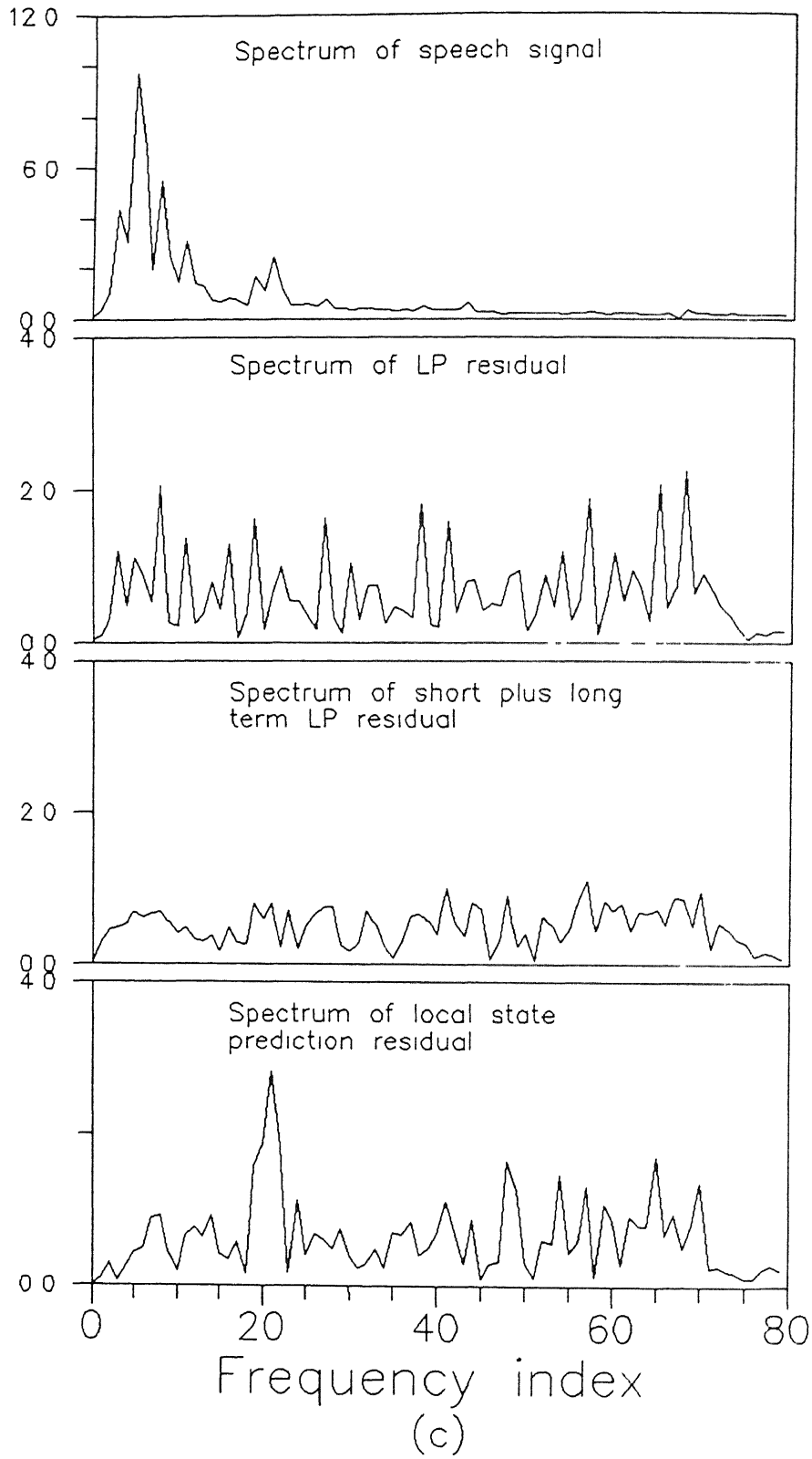
Fig. 5.8: Comparative plots of the speech signal, short term LP residual, short term plus long term LP residual and local state prediction residual in terms of (a) time series, (b) autocorrelation function, and, (c) DFT spectrum



Time series plots – 10^{th} order LP used to remove short term correlations from the speech signal The long term LP uses 3 coefficients The LSP parameters are $N_f = 320$, $N_L = 40$ and $d = 10$ The SNR based on 160 samples for the short term LP residual, short term plus long term LP residual and LSP residual are 21 88 dB, 26 61 dB and 21 41 dB respectively

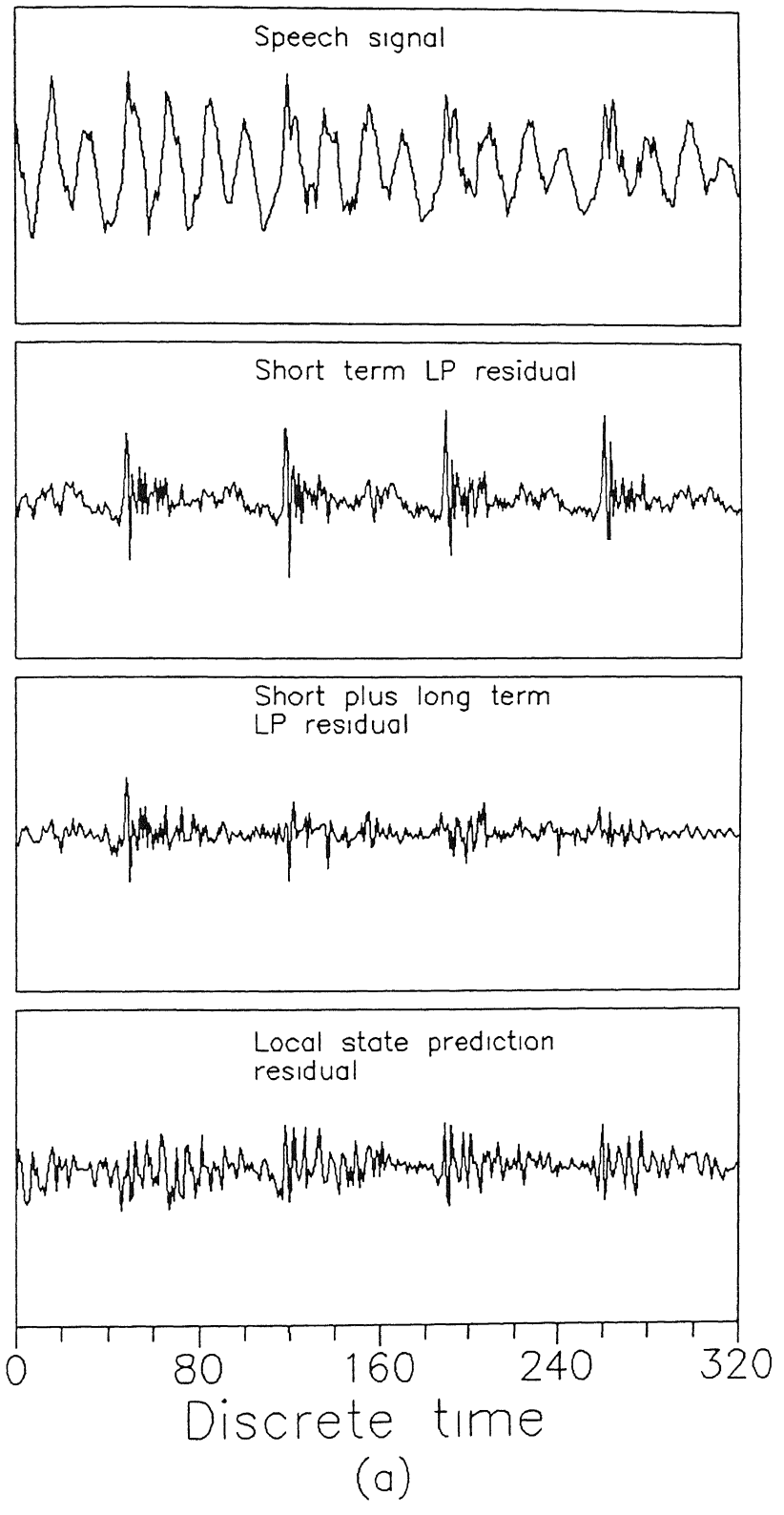


Autocorrelation function plots corresponding to the 4 time series plots of part (a)

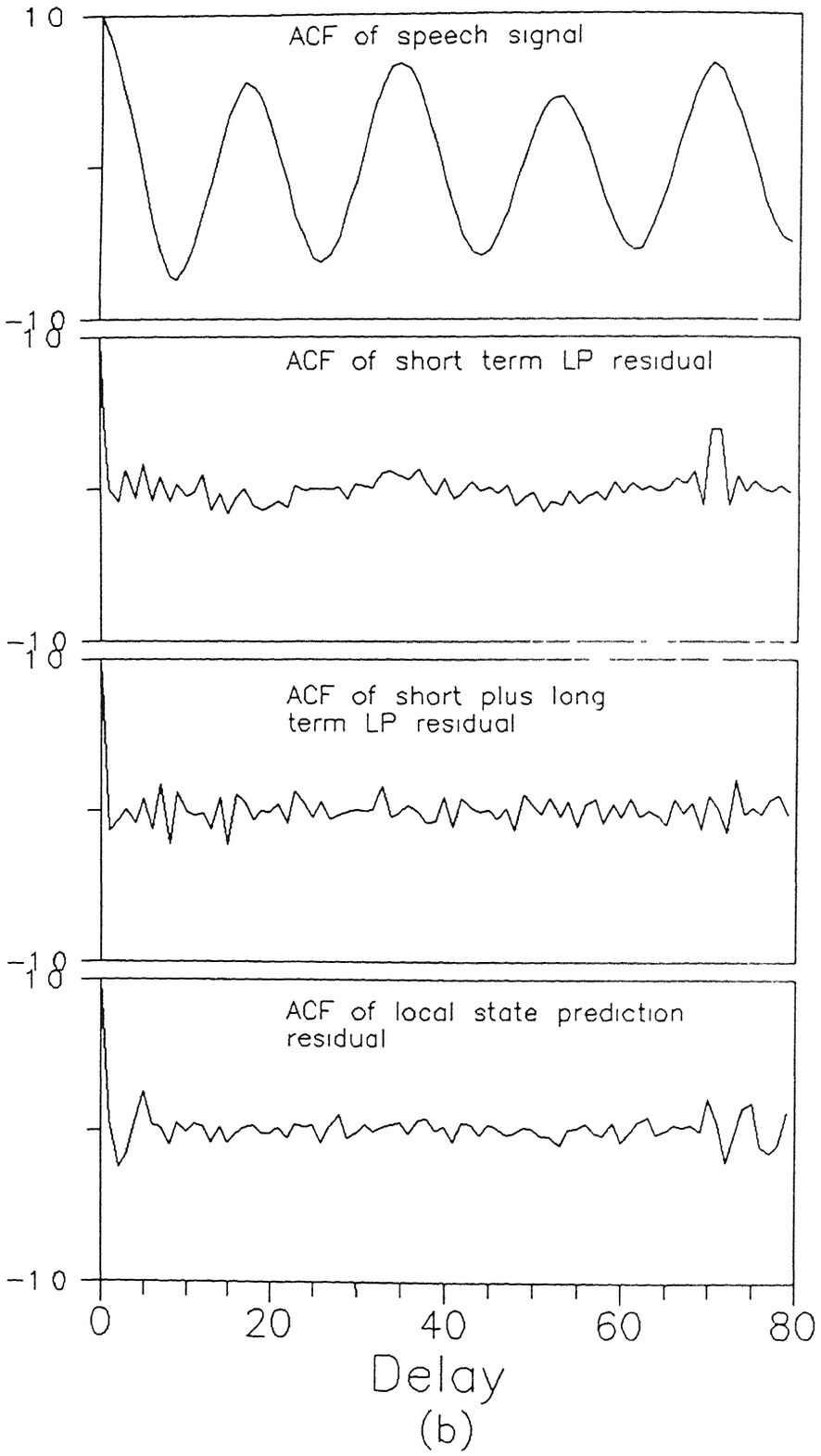


Spectrum plots corresponding to the 4 time series plots of part (a) computed using 160 point DFT Note the relative magnitudes of the plots

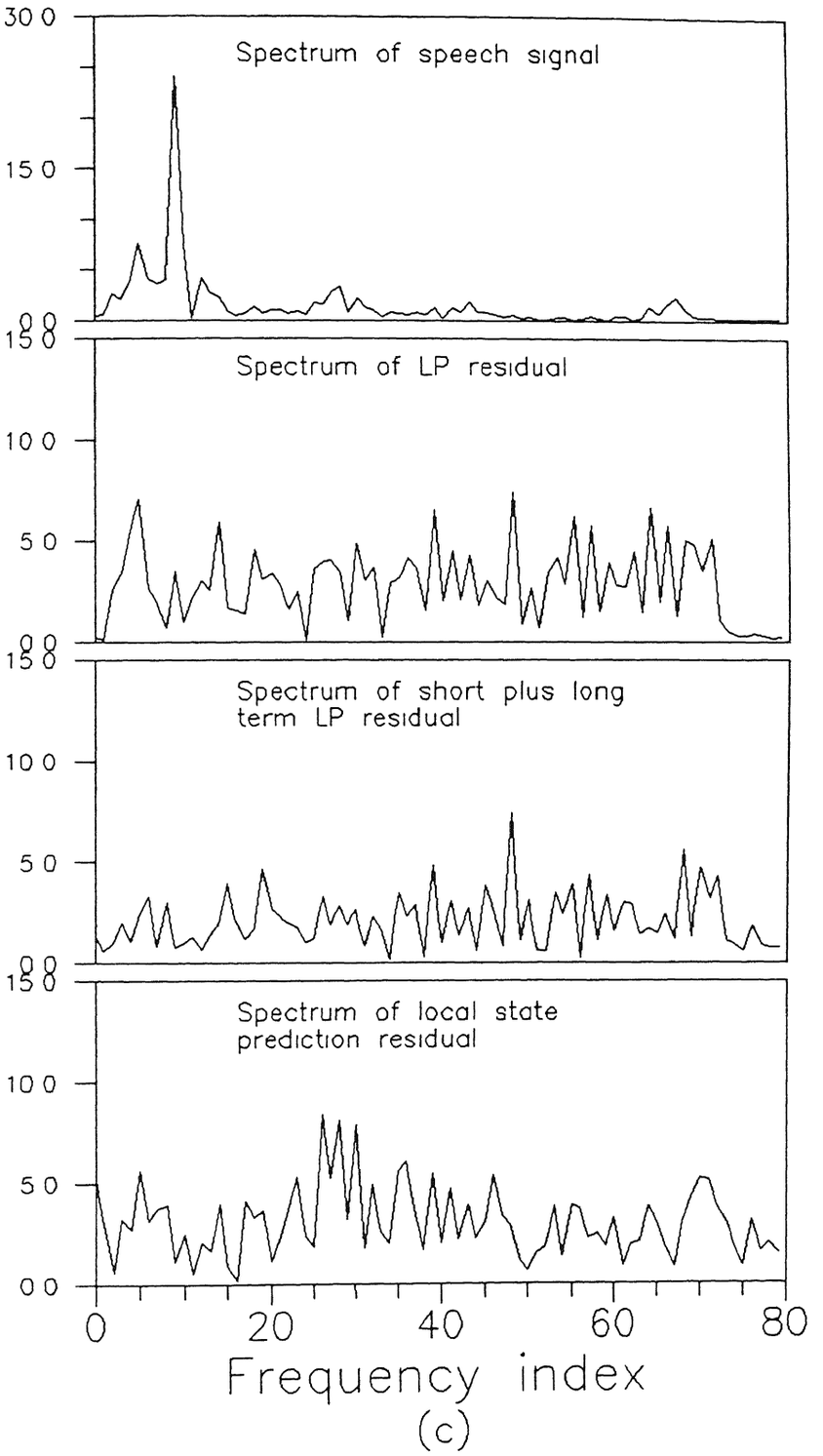
Fig. 5 9: Comparative plots of the speech signal, short term LP residual, short term plus long term LP residual and local state prediction residual in terms of (a) time series, (b) autocorrelation function, and, (c) DFT spectrum



Time series plots – 10^{th} order LP used to remove short term correlations from the speech signal The long term LP uses 3 coefficients The LSP parameters are N_f = 320, N_L = 40 and d = 10 The SNR based on 160 samples for the short term LP residual, short term plus long term LP residual and LSP residual are 12 06 dB, 16 51 dB and 13 04 dB respectively

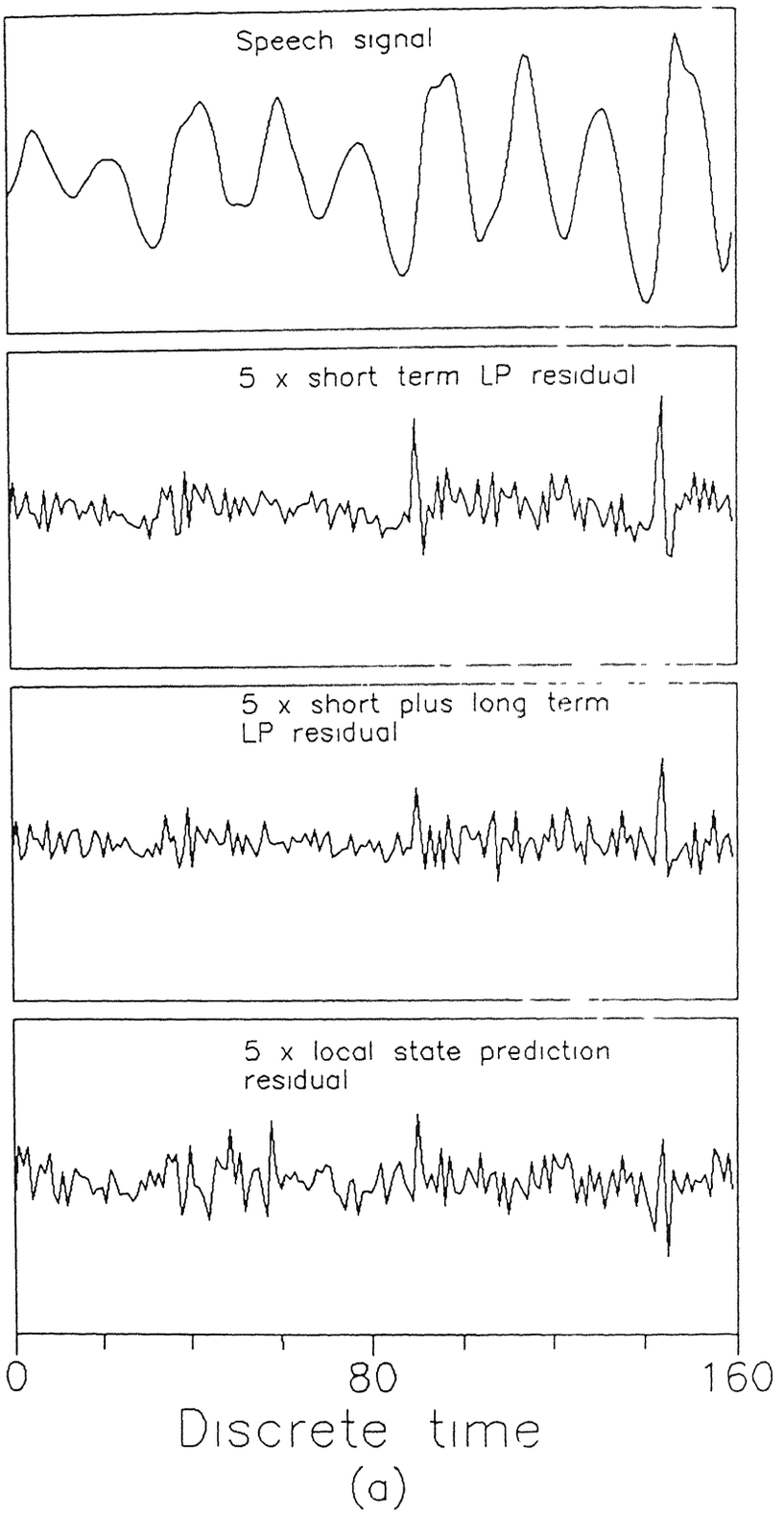


Autocorrelation function plots corresponding to the 4 time series plots of part (a)

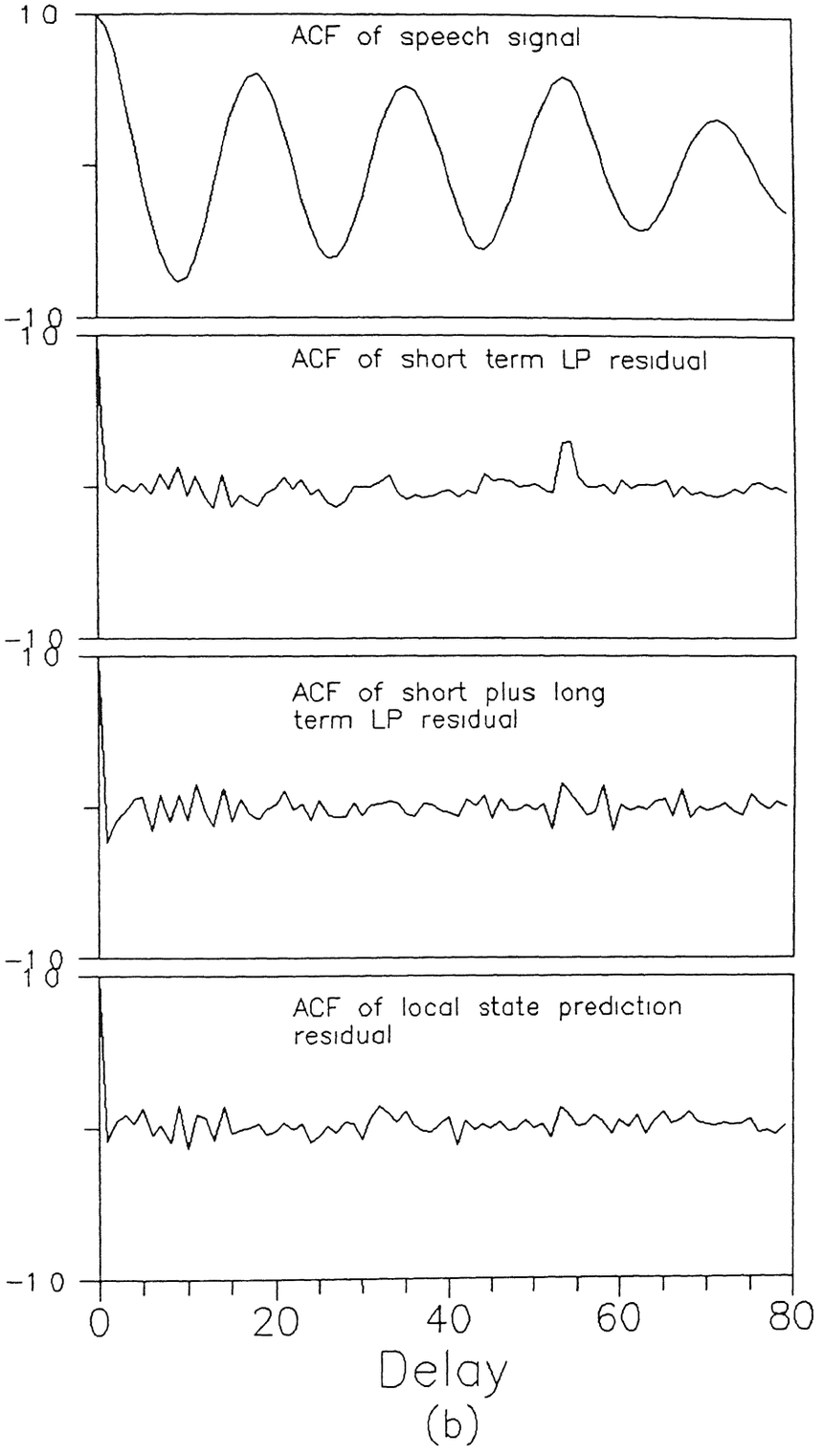


Spectrum plots corresponding to the 4 time series plots of part (a) computed using 160 point DFT Note the relative magnitudes of the plots

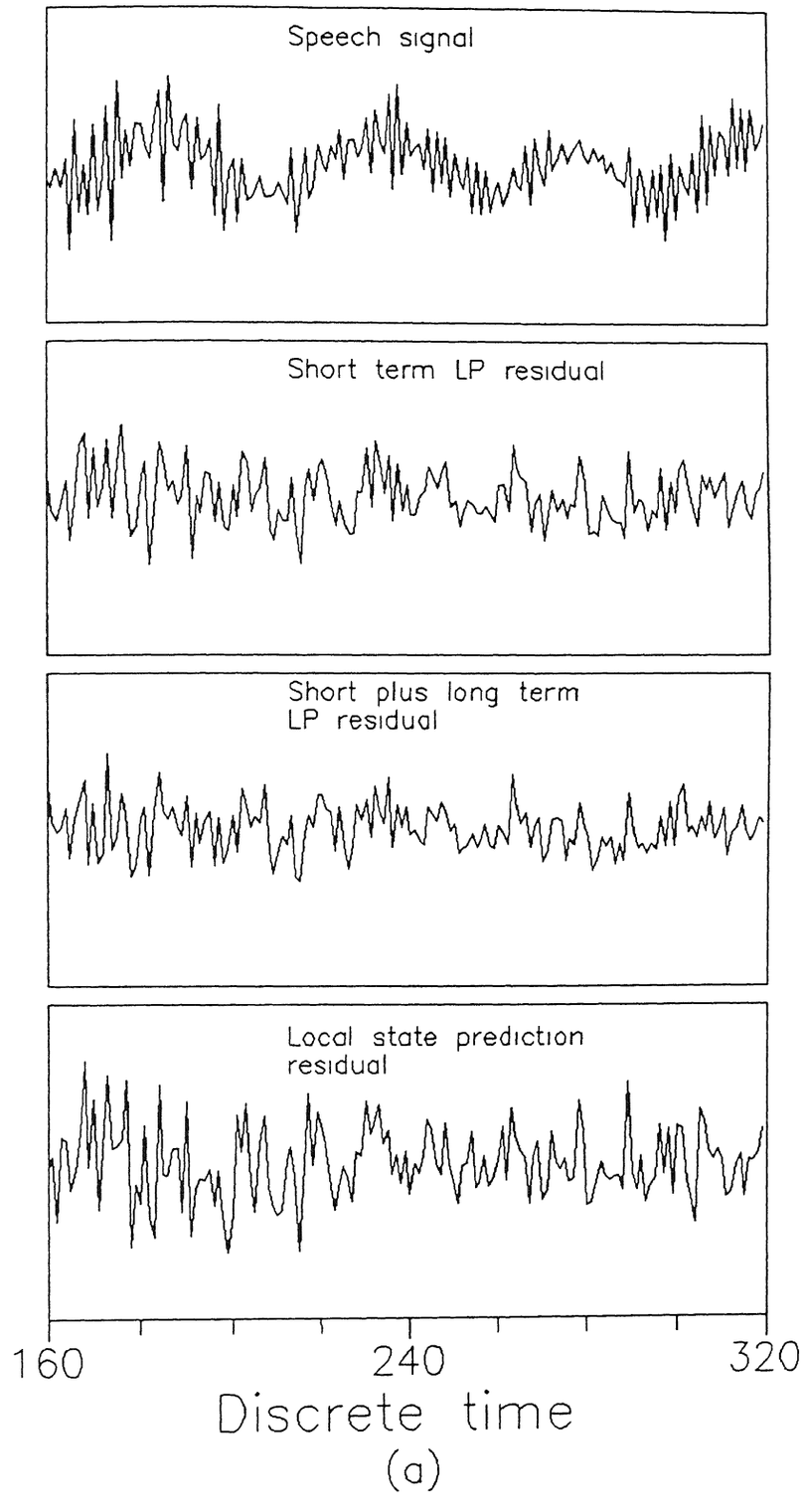
Fig 5 10: Comparative plots of the speech signal, short term LP residual, short term plus long term LP residual and local state prediction residual in terms of (a) time series, (b) autocorrelation function, and, (c) DFT spectrum



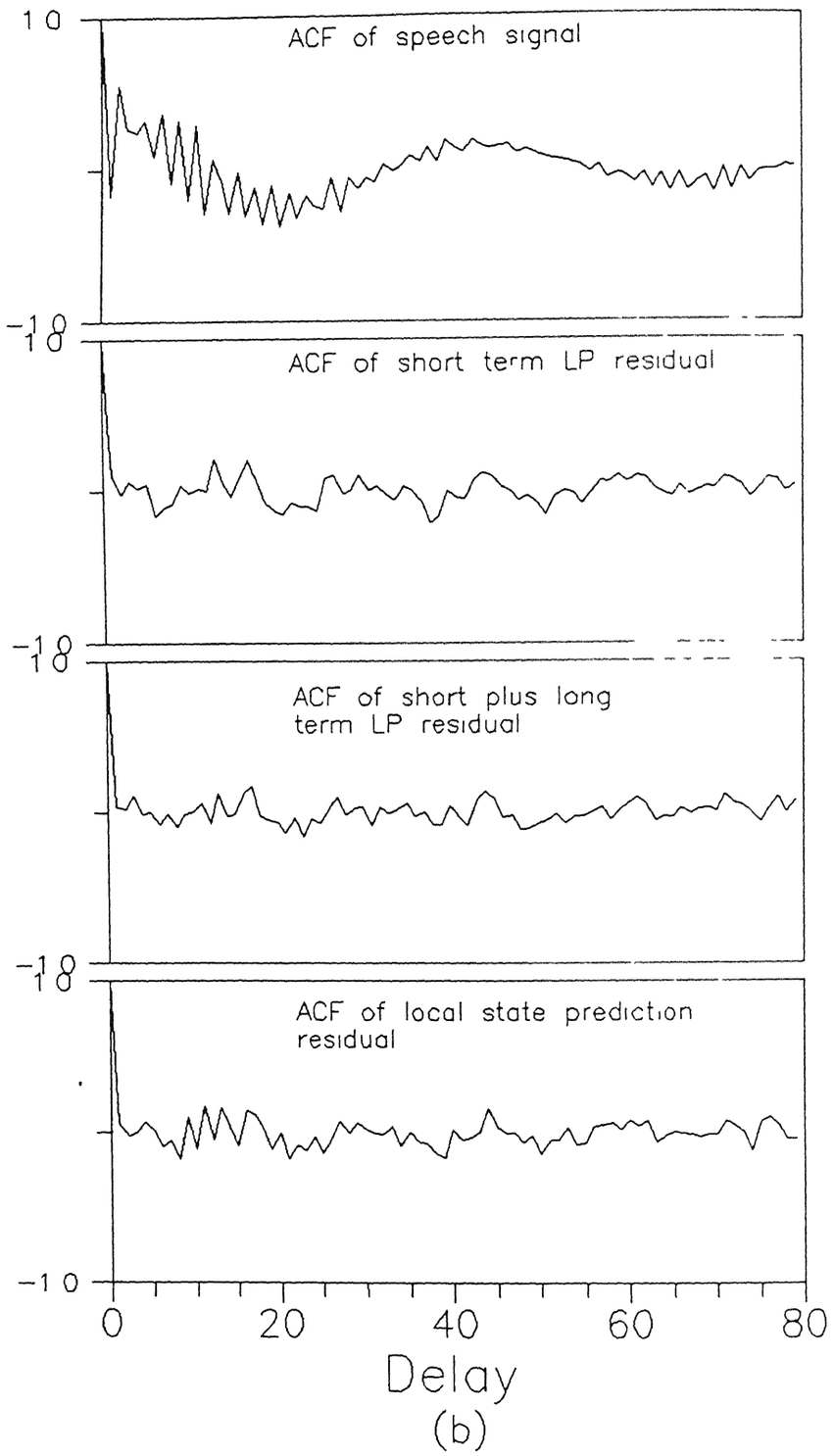
Time series plots – 10^{th} order LP used to remove short term correlations from the speech signal The long term LP uses 3 coefficients The LSP parameters are $N_f = 320$, $N_L = 40$ and $d = 10$ The SNR based on 160 samples for the short term LP residual, short term plus long term LP residual and LSP residual are 27 02 dB, 30 14 dB and 27 26 dB respectively



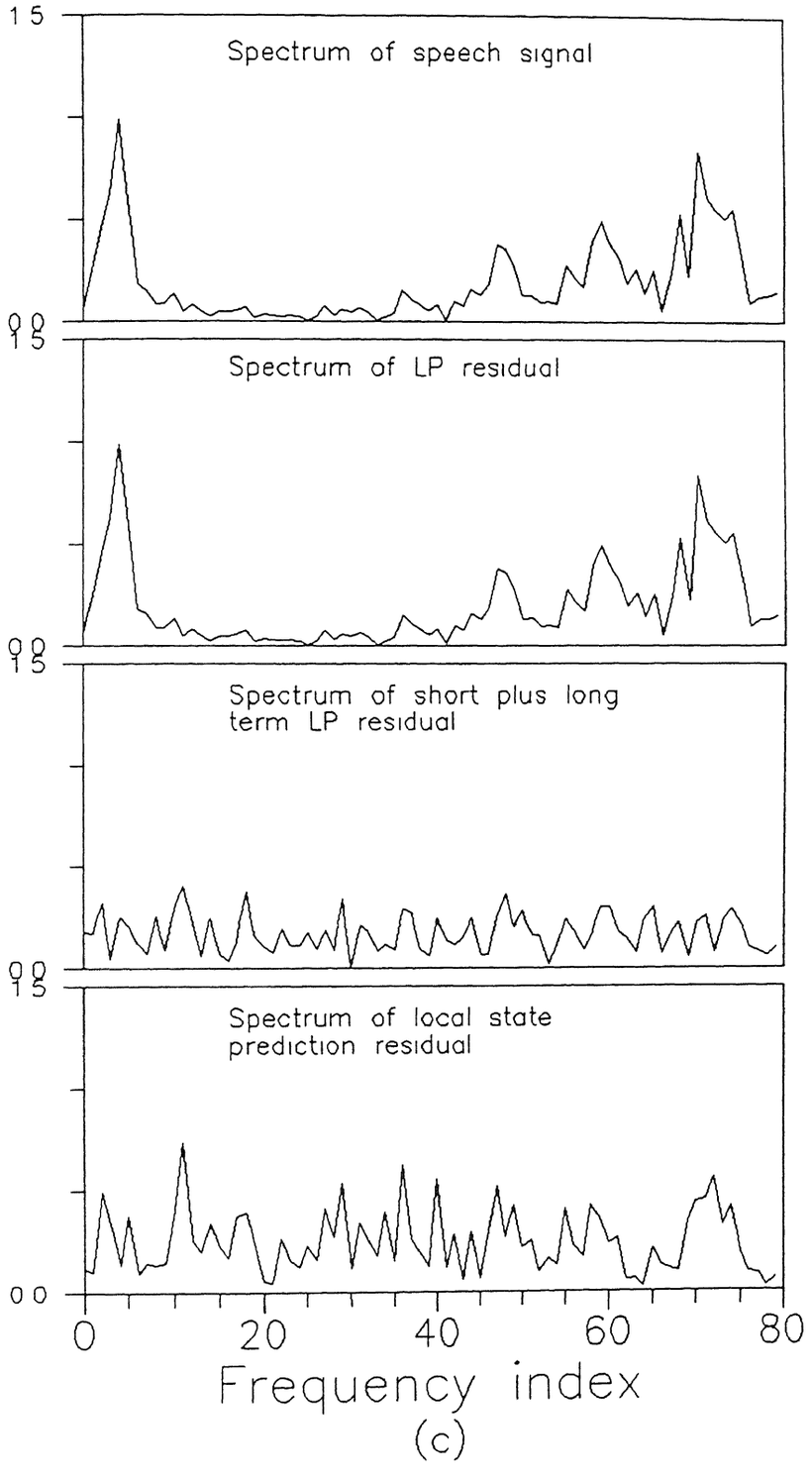
Autocorrelation function plots corresponding to the 4 time series plots of part (a)



Time series plots – 10^{th} order LP used to remove short term correlations from the speech signal The long term LP uses 3 coefficients The LSP parameters are $N_f = 320$, $N_L = 40$ and $d = 10$ The SNR based on 160 samples for the short term LP residual, short term plus long term LP residual and LSP residual are 8 68 dB, 9 77 dB and 6 51 dB respectively



Autocorrelation function plots corresponding to the 4 time series plots of part (a)



Spectrum plots corresponding to the 4 time series plots of part (a) computed using 160 point DFT Note the relative magnitudes of the plots

Fig 5.12 Comparative plots of the speech signal, short term LP residual, short term plus long term LP residual and local state prediction residual in terms of (a) time series, (b) autocorrelation function, and, (c) DFT spectrum

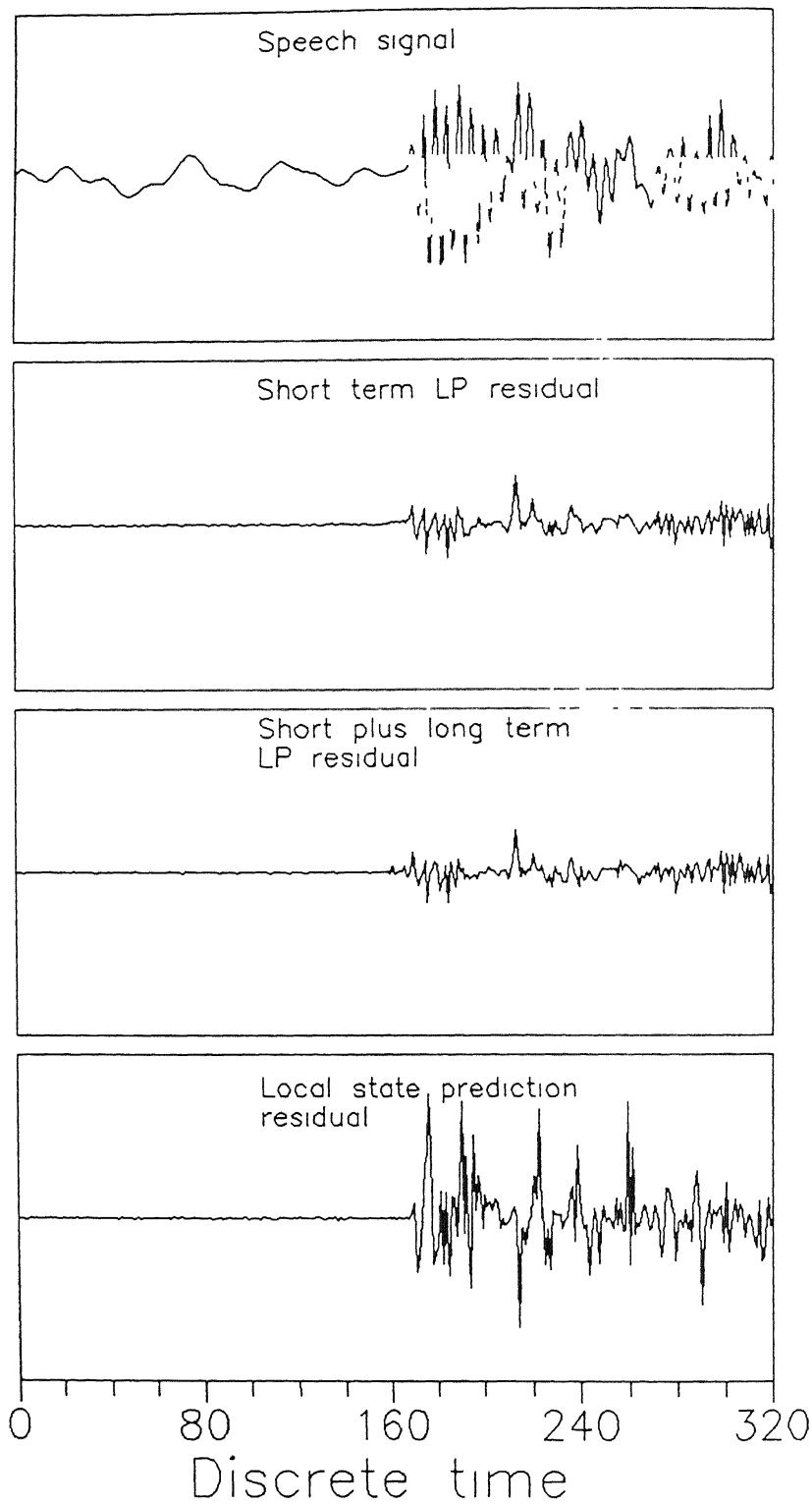


Fig. 5.13: Time series plots to show the inadequacy of the LSP scheme to track sudden changes in the data compared to the forward block adaptive case of the LP filters The SNRs for the two 160 sample frames are (i) 25 22 and 10 89 dB, (ii) 26 18 and 10 53 dB and (iii) 22 67 and 0 84 dB for (i) short term LP residual, (ii) short term plus long term LP residual and (iii) local state prediction residual respectively

signal amplitude in the form of a voiced utterance in the second frame which continues in subsequent frames The LSP fails to predict the transient in the second frame and the SNR picks up only in subsequent frames

Frame Number	1	2	3	4	5
Short term LP residual	25 22	10 39	11 44	22 00	23 52
Short plus long term LP residual	26 18	11 03	12 38	23 08	26 39
LSP residual	22 67	0 84	6 69	18 29	24 63

Table 6.1: SNR values for successive 160 sample frames to illustrate the inadequacy of the LSP scheme to track sudden changes in signal characteristics.

Computational Complexity Considerations: A local state prediction is significantly more expensive in terms of computational complexity compared to the usual linear prediction To predict one sample using LSP, the major sources of computational complexity are in the following three operations

(1) To find N_L nearest neighbours from N_f trajectory points one has to compute the $(N_f - 1)$ distances from the target vector in d-dimensional space

(2) To fit an optimal local *linear* predictor between the N_L nearest neighbours and their future points, one has to compute a $(d+1) \times (d+1)$ covariance matrix that requires $O(N_L d^2)$ multiplications Note that a simplification to $O(N_L d)$ multiplications as in the case of linear prediction is not possible

(3) To solve for the $(d+1)$ coefficients of the predictor, an additional $O(d^3)$ multiplications are required with the Cholesky decomposition method

One can compare the above with the computational complexity of a linear prediction scheme which requires $O(N_f d)$ multiplications to compute the entries of the covariance matrix and $O(d^3)$ multiplications to get the d optimal LP coefficients These computations are required to predict *all* the N_f samples within an analysis frame in the forward adaptive case

5.4 Recent Studies in Local Methods for Speech Prediction and Coding

Some researchers have reported the study of local methods for speech prediction and coding in the recent past We will discuss the salient features of these schemes in terms of the terminology used in the previous sections

One of the first methods in this category is the Pattern Search Predictor (PSP) of Bogner and Li [18] This predictor is similar to the LSP except for an additional signal matching scheme to scale all reconstructed vectors by appropriate norms to account for variations in signal amplitude Only a single neighbour is chosen (i e, $N_L = 1$) from an analysis frame of length N_f, and the prediction of the target vector is taken to be the future sample of the nearest neighbour They have incorporated a PSP in a standard CCITT 32 kb/s ADPCM coder and studied its performance

A preliminary version of the LSP, termed as the Compromised Overlapping Neighbourhood (CON) – Local Approximation Technique was studied and reported by us in [90], [92], [93] In this approximation technique, the prediction is performed *within* the analysis frame While this scheme gives better prediction properties compared to LSP, it is not suitable for use in a speech coder for obvious reasons

In another study due to Townshend [143], [144], a prediction scheme similar to LSP is used on the LP residual in an ADPCM speech coder The local predictor, however, finds a local neighbourhood from a *fixed* analysis frame of LP residuals corresponding to a relatively large frame length of 30 s of speech

In a paper due to Wang *et al*, [119], a nonlinear predictor for speech based on vector quantization techniques is studied In this scheme, two vector codebooks are designed corresponding to voiced and unvoiced speech The training vectors used in the design of the codebooks are of the form given in eq (5 1) Two scalar codebooks are also designed corresponding to approximately optimal predictions of the codevectors of the respective codebooks A scalar prediction of a *target* vector is done by making a voiced/unvoiced decision for the vector, finding the nearest codevector in the appropriate codebook and outputting the scalar value corresponding to that codevector index in the associated scalar codebook. This method can also be

regarded as a local prediction scheme Compared to LSP, this prediction scheme uses a fixed analysis frame of very large length (corresponding to the training set size), a fixed neighbourhood partition and a zeroth order local predictor (corresponding to the simple averaging of the future points of the training vectors in each local neighbourhood)

A "nonlinear oscillator model" for the reproduction of speech waveform has been reported very recently by Kubin and Kleijn [89] This model also essentially employs a local prediction scheme It uses an adaptive analysis frame However, its differences with LSP lie mainly in the use of a more general reconstructed vector compared to eq (5 1), and in the use of a zeroth order local predictor

Singer *et al* [129] have interpreted the problem of local state prediction as a codebook prediction problem In this interpretation, the reconstructed trajectory vectors within an analysis frame and their future points, together form a codebook pair Given a target vector whose future sample is to be predicted, the same steps as enunciated above for LSP are followed

5.5 Structure and Performance of a Vector Excited Local State Prediction (VELSP) Coder

A speech coding scheme which incorporates a local state predictor must be tailored to take advantage of the features of the predictor Two observations regarding a LSP are as follows (i) it simulates the function of both short and long term predictors to make predictions, and (ii) it is more suitable for use in a backward adaptive form where a prediction is made ahead of an analysis frame rather than within it as in forward adaptive prediction Based on the prediction performance of a LSP and the above observations, we study a basic speech coding scheme using LSP This coder is suitable for further development in the medium bit rate range of 4 8 – 16 kb/s range and can theoretically provide low to medium coding delay The structure of the coder and the decoder are shown in figs 5 14 and 5 15 respectively This is an analysis – by – synthesis coding scheme and is similar to CELP in principle We tentatively name the coder as a Vector Excited Local State Prediction (VELSP) coder

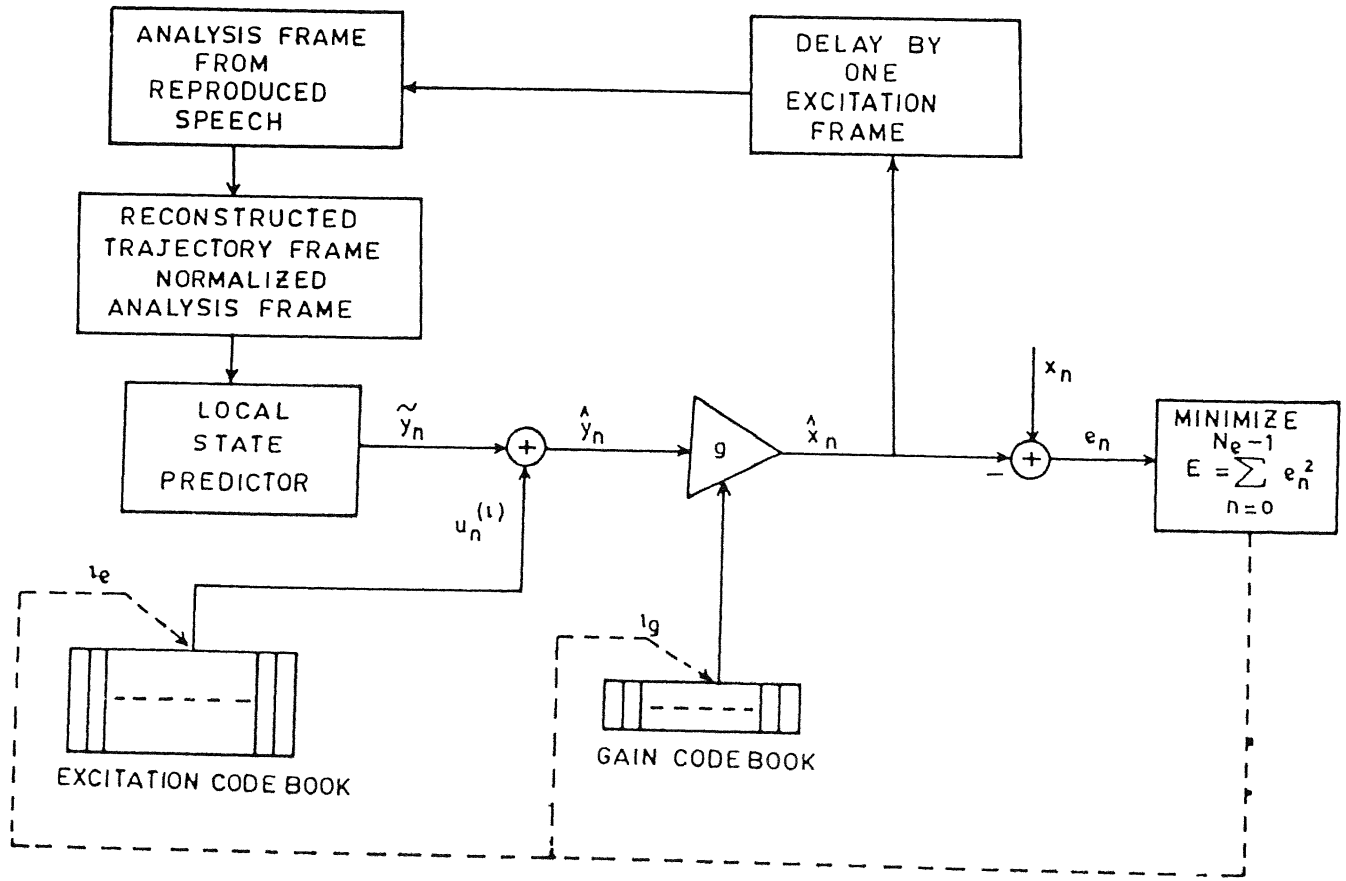


Fig 5.14. Basic structure of a Vector Excited Local State Prediction (VELSP) coder

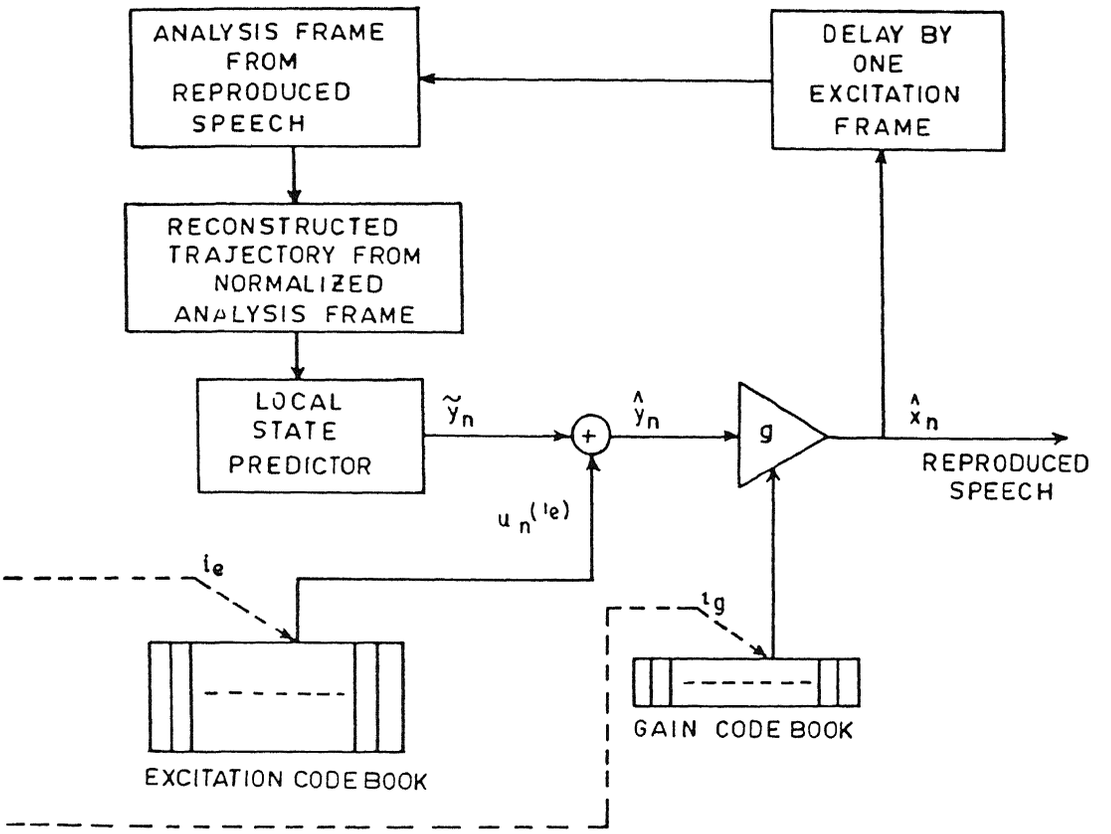


Fig. 5.15: Basic structure of a Vector Excited Local State Prediction (VELSP) decoder

There are two frame lengths involved in this coding scheme One is the *analysis frame* which consists of *previous* reproduced speech based on which LSP is performed and the second is the *excitation frame* whose length denotes the excitation sequence length and the length for which prediction is performed based on one analysis frame

The crucial difference of a VELSP from a CELP is in the use of a nonlinear predictor rather than a linear prediction filter Consequent to this, it is not possible to break up the excitation vector into a "gain" factor and a "shape" vector and compute the indices from the respective codebooks in the usual way One possible method of overcoming this difficulty is to use an excitation vector codebook which incorporates the gain factor However, given the dynamic range of speech, a large size codebook would be required which makes an optimal excitation vector search impractical We attempt to overcome this problem by normalizing the reproduced speech used in the analysis frame, performing LSP based on the normalized frame and doing a gain multiplication of the "normalized" predicted frame in the end to get a frame length of reproduced speech Thus, the steps involved in getting an excitation frame length of reproduced speech are as follows

(1) Normalize the analysis frame length of immediately previous reproduced speech,

(2) Corresponding to an index i of the excitation codebook, do the following

   (i) Get one excitation frame length of "normalized" reproduced speech,

   (ii) Get the corresponding optimal gain factor by minimizing the m s e between this excitation frame length of "normalized" reproduced speech and the original speech The *reproduced speech* is given by the gain multiplied by the "normalized" reproduced speech

(3) Repeat step (2) for all indices i of the excitation codebook The coder transmits that index i_e and the corresponding quantized gain index i_g which minimize the m s e between the original and reproduced speech of one excitation frame length

In the following, we will discuss the analysis steps of a basic VELSP coder, the design of excitation vector and gain codebooks and the performance characteristics and scope for improvement of the coder

A. Analysis Steps of a VELSP Coder

Let us consider the analysis steps required to obtain the parameters for transmission corresponding to one excitation frame length N_e of input speech, x_n, $n = 0, 1, \quad , N_e - 1$ We will need (i) an "analysis frame" of *reproduced* speech, $\hat{x}_n$, $n = -N_f - d + 1, \quad , -1$, where d is the embedding dimension and N_f is the *analysis frame length* corresponding to the length of the reconstructed trajectory in d- dimensional space, (ii) an "excitation codebook" of size N_c consisting of excitation sequences $u_n^{(i)}$, $n = 0, \quad , N_e - 1$, $i = 1, \quad , N_c$, and (iii) a scalar "gain codebook" of size N_g, consisting of gain values $g^{(i)}$, $i = 1, \quad , N_g$ We will discuss about the design of the two codebooks after an understanding of their functions Given the above, the analysis steps are as follows

STEP 1: Obtain the *normalized* trajectory of reproduced speech $\hat{\mathbf{y}}_n^d$, $n = -N_f, \quad , -1$, where

$$\hat{\mathbf{y}}_n^d = [\hat{y}_{n-d+1}\hat{y}_{n-d+2} \quad \hat{y}_{n-1}\hat{y}_n]^T, \tag{5 9}$$

$$\hat{y}_n = \frac{\hat{x}_n}{G_{rms}}, \quad n = -N_f - d + 1, \quad , -1 \tag{5 10}$$

and

$$G_{rms} = \left(\frac{1}{N_f + d - 1} \sum_{n=-N_f-d+1}^{-1} \hat{x}_n^2 \right)^{1/2} \tag{5 11}$$

STEP 2: Corresponding to an excitation codebook sequence index i,

(i) Obtain "normalized" reproduced speech $\hat{y}_n$, $n = 0, 1, \quad , N_e - 1$ according to the following

$$\hat{\mathbf{y}}_n^d = \mathbf{g}\left(\hat{\mathbf{y}}_{n-1}^d\right) + u_n^{(i)}\mathbf{h} \tag{5 12a}$$

$$\hat{y}_n = \mathbf{h}^T\hat{\mathbf{y}}_n^d, \quad n = 0, 1, \quad , N_e - 1 \tag{5 12b}$$

where,

$$\mathbf{h} = [0 \quad 01]^T \tag{5 13a}$$

$$\mathbf{g}(\hat{\mathbf{y}}_{n-1}^d) = [\hat{y}_{n-d+1}\hat{y}_{n-d+2} \quad \hat{y}_{n-1}\tilde{y}_n]^T \tag{5 13b}$$

and

$$\tilde{y}_n = f(\hat{\mathbf{y}}_{n-1}^d) \tag{5 13c}$$

Here, $f(\hat{\mathbf{y}}_{n-1}^{d})$ is a local state predictor whose form is chosen to be a linear model plus a constant term, eq (5 4), in the coder studied by us

(ii) Get the corresponding optimal gain factor

$$g = \frac{\sum_{n=0}^{N_e-1} r_n \hat{y}_n}{\sum_{n=0}^{N_e-1} \hat{y}_n^2} \tag{5 14}$$

The above equation results by minimizing the m s e E with respect to g, where

$$E = \frac{1}{N_e} \sum_{n=0}^{N_e-1} (r_n - g\hat{y}_n)^2 \tag{5 15}$$

Also,

$$\hat{x} = \hat{g}\hat{y}_n, \quad n = 0, 1, \quad , N_e - 1 \tag{5 16}$$

is the reproduced speech corresponding to the excitation index i and the *quantized* gain factor $\hat{g}$

STEP 3: Repeat step 2 for excitation indices $i = 1, \quad , N_c$ Choose the index i_e and the corresponding quantized gain index i_g which minimizes

$$E_r = \frac{1}{N_e} \sum_{n=0}^{N_e-1} (x_n - \hat{r}_n)^2 \tag{5 17}$$

The two indices i_e and i_g are transmitted per excitation frame to the decoder The decoder also has the analysis frame of reproduced speech $\hat{x}_n$, $n = -N_f - d + 1, \quad , -1$ It executes steps 1(i) and eq (5 16) to get the reproduced speech, $\hat{x}_n$, $n = 0, 1, \quad , N_e - 1$

The above analysis shows that one needs to store input speech of length N_e before any processing for the frame can begin The theoretical minimum coding delay will therefore be given by $3N_e$ samples or $\frac{3N_e}{f}$ s where f is the sampling frequency of speech

B. Design of Excitation and Gain Codebooks

We give herein an approximate procedure for the design of the excitation and gain codebooks required by the VELSP coder and decoder Given a sequence of speech data, the corresponding vector and scalar training sequences are found from the following steps

STEP 1: For one *frame* of speech data, x_n, $n = -N_f - d + 1, \quad , -1, 0, 1, \quad , N_e - 1,$

(i) Get a "normalized" frame y_n, $n = -N_f - d + 1, \quad , N_e - 1$, where

$$y_n = \frac{x_n}{\hat{G}_{rms}} \tag{5 18a}$$

and

$$\hat{G}_{rms} = \left[\frac{1}{N_f + d + N_e - 1} \sum_{n=-N_f-d+1}^{N_e-1} x_n^2 \right]^{1/2} \tag{5 18b}$$

(ii) Predict $\hat{y}_0$ using LSP on the analysis frame y_n, $n = -N_f - d + 1, \quad , -1$ using the procedure given above

STEP 2: For $n = 1, \quad , N_e - 1$, do the following

(i) Shift the analysis frame by one sample to the future,

(ii) Predict $\hat{y}_n$

This gives one training sequence $\hat{y}_n$, $n = 0, 1, \quad , N_e - 1$ for the excitation codebook.

STEP 3: A training point g for the gain codebook is obtained using eq (5 14) which results from the minimization of the m s.e. between x_n and y_n, $n = 0, 1, \quad , N_e - 1$

From successive frames of speech, one can get the training sets using steps 1–3 for the design of the two codebooks The gain and excitation codebooks can then be designed using the Lloyd and the generalized Lloyd algorithms respectively [52]

C. Performance of a Fully Quantized VELSP Coder

We have implemented and done preliminary performance study of a basic VELSP coder at three operation rates of 8 0 kb/s, 6 5 kb/s and 5 2 kb/s. The coder parameters used for each bit rate of operation are shown in Table 5 2 It is seen that the theoretical minimum coding delay at 5 2, 6 5 and 8 0 kb/s operating rates is equal to 7 5 ms, 6 0 ms and 4 875 ms respectively

Codebook Design: The excitation and gain codebooks were designed using the steps given above for each bit rate of operation The two training sets were obtained from the 4 phoneme – specific sentence utterances of speech database 2 (Appendix B) by 3 male and 3 female speakers The total duration of speech signal used corresponding

Bit Rate, kb/s	5.2	6.5	8.0
Analysis frame length, N_f	160	160	169
Embedding dimension, d	10	10	10
Local neighbourhood size, N_l	40	40	40
Excitation frame length, N_e	20	16	13
Excitation codebook size, N_c	256	256	256
Gain codebook size, N_g	32	32	32

Table 6.2: VELSP coder parameters at three bit rates of operation

to these sentence utterances is 50 s which provides 400000 samples at 8 kb/s sampling rate The training ratios for the excitation codebook at the 3 bit rates of 5 2, 6 5 and 8 0 kb/s are 76 1, 95 0 and 116 8 respectively The training ratios for the corresponding gain codebooks are 609 1, 760 3 and 934 8 respectively The initial codebook for the excitation sequences was designed using the method of pruning [52] Subsequently, iterative refinement was performed using the generalized Lloyd algorithm The gain codebook was designed from its training set using the Lloyd algorithm for empirical data

Performance Study: The performance of the coder was studied using the 4 sentences (a)–(d) of database 2 (Appendix B) spoken by one male and one female This test data is different from that used for the design of the codebooks The segmental SNR performance at the 3 bit rates is shown in fig 5 16 for the 4 phoneme – specific sentences (a)–(d) and the overall database (dashed line) The segment length used for the computation of the segmental SNR is 160 samples for 5 2 and 6 5 kb/s rates and 169 samples for the 8 0 kb/s rate As expected, this measure of objective performance degrades with the lowering of the bit rate of operation

The reproduced speech has perceptible noise and becomes poor at the 5 2 kb/s rate It must be mentioned that this performance is for the skeletal structure of the VELSP coder/decoder in which we have not incorporated any standard performance enhancing schemes.

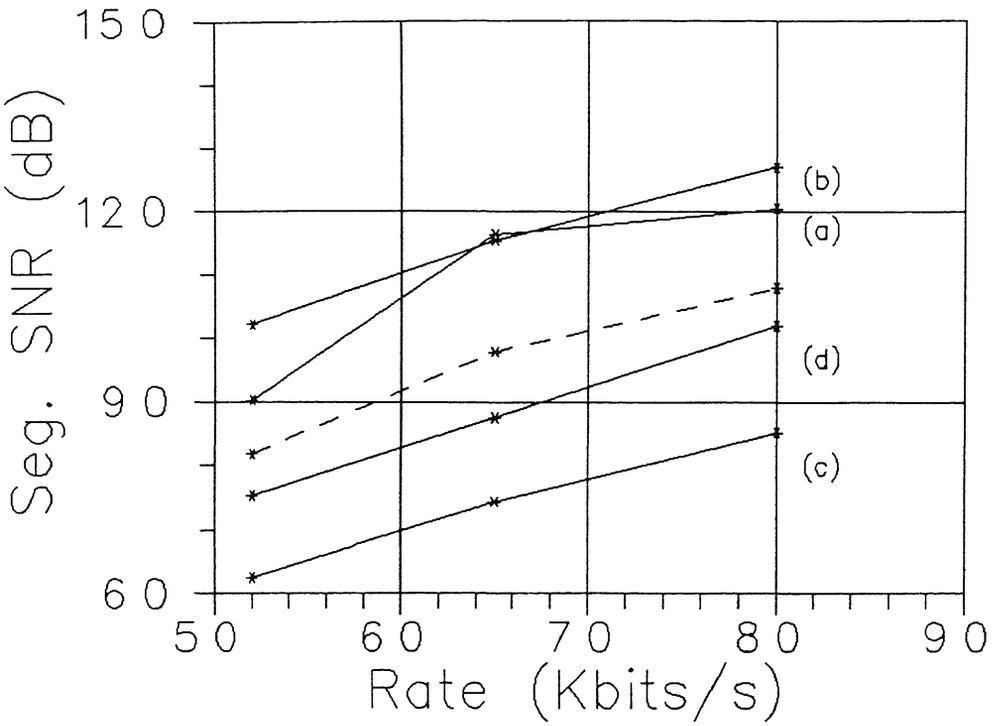


Fig. 5.16: Segmental SNR performance of a VELSP coding scheme at three bit rates of 5 2, 6 5 and 8 0 Kbits/s. Graphs (a)–(d) are based on the phoneme specific sentences (a)–(d) respectively (speech database 2, Appendix B) spoken by one male and one female The dashed line gives the performances for the overall database

We also compared the objective and subjective performance of the fully quantized VELSP coder with *unquantized* forward adaptive and backward adaptive CELP coders without perceptual weighting The forward adaptive CELP coding scheme that was implemented is discussed in chapter 4 The backward adaptive CELP that was implemented is a block adaptive scheme in which the short term LP parameters for an excitation were computed from an analysis frame length of immediately previous reproduced speech values The parameters of the CELP coders were chosen such that when quantized, they would operate at approximately 5 2, 6 5 and 8 0 kb/s The following qualitative statements can be made about the comparisons

(1) The segmental SNR performance of unquantized forward adaptive CELP coders is 1 3 to 1 4 dB better than quantized VELSP coders at the three rates Similarly, the segmental SNR of *unquantized* backward adaptive CELP coders is 1 0 to 1 1 dB more than quantized VELSP coders at the three rates Kroon and Atal have noted that the segmental SNR of unquantized CELP is approximately 2 dB more than that of a fully quantized one [84]

(2) The subjective performance of a quantized VELSP coder is somewhat better than a backward adaptive CELP coder

Directions for Further Study: The above preliminary investigation of a VELSP coding scheme suggests that it can be the subject of a more detailed investigation There are two major directions in which efforts are required These are

(1) Improvement in subjective quality performance Instead of a simple m s e minimization, a perceptually weighted minimization criterion, as in CELP, can be incorporated Further effort is required to reduce the perceptible noise in the reproduced speech by doing a detailed study of the characteristics of the noise and incorporating, for example, an adaptive postfiltering scheme

(2) Reduction in computational complexity of the coder and decoder The major sources of computational complexity for *one* local state prediction are as given section 5 3. For a VELSP coding scheme, in order to derive a gain and an excitation index corresponding to N_e speech samples, a full complexity coder will need to perform $N_e \times N_c$ local state predictions where N_c is the size of the excitation codebook

Likewise the decoder has to perform N_e local state predictions in order to get the corresponding N_e reproduced speech samples. While there are some obvious methods to reduce the computational complexity marginally, a detailed effort is required to investigate the possibility of bringing it down significantly

In conclusion, we state that the above paradigm for low delay speech coding can be gainfully used when positive outcomes are obtained with regard to the above two directions of further investigation

Chapter 6

The Rate Distortion Function and Computation of a Lower Bound

Rate distortion theory is the discipline that deals with issues of signal compression from an information theoretic viewpoint The basic problem here can be stated as follows given a source distribution and a distortion measure, what is the minimum rate at which information about the source can be reproduced with a specified expected distortion? The foundation of rate distortion theory was laid by Shannon in 1959 in his paper, "Coding theorems for a discrete source with a fidelity criterion," where he defined the rate distortion function of an information source [127]. Prior to this, he had already introduced the concept of rate distortion in his celebrated paper, "A mathematical theory of communication," in 1948 [126] Specifically, the rate distortion function, $R(D)$, gives the minimum rate R at which information about a source can be transmitted subject to the constraint that it can be reproduced with an average distortion D. The rate at which a source produces information subject to the requirement of lossless or perfect reproduction is the entropy of the source Rate distortion function can therefore be looked upon as a generalization of the concept of entropy It may be worthwhile to keep in mind the basic communication system block diagram of fig 1 1 for the deliberations in this chapter

According to the information transmission theorem, which is a generalization of the channel coding theorem, it is impossible to obtain an average distortion D

or less unless $R(D) < C$, where C is the capacity of the transmission channel For memoryless sources, one can usually compute $R(D)$ analytically if the source p d f is known It can also be computed numerically from source output realizations using Arimoto and Blahut's algorithm [16], [17] For sources with memory, one can capitalize on the inherent statistical dependencies to further reduce the minimum rate needed to achieve a specified average distortion Even for source outputs described by independent random variables, joint descriptions are more efficient compared to individual descriptions This is because rectangular grid points which arise naturally in independent descriptions do not fill the space as efficiently as lattice grid points in n-space Thus, it is meaningful to compute the rate distortion function from joint distributions of the source output process However, this is a difficult task Analytically, $R(D)$ is known only for a few joint p d f's such as the joint Gaussian density function and for specific distortion criteria [12] One can attempt to compute it from source output realizations using Arimoto and Blahut's algorithm but the computational effort and data length requirement increase exponentially as one considers statistical dependencies extending to successively larger time frames

In this chapter, we consider the computation of a lower bound $R^L(D)$ of $R(D)$ for stationary ergodic sources with memory Specifically, $R^L(D) = R_1(D) - \Delta$, where $\Delta = H(X) - H$ for discrete alphabet sources and $\Delta = h(X) - h$ for continuous alphabet sources [12], [151] Here, $R_1(D)$ is the first order rate distortion function and $H(X)$ {h(X)} is the entropy {differential entropy} corresponding to the marginal probability {marginal probability density} function of the source output and H { h} is the entropy rate {differential entropy rate} with respect to the joint probability {joint probability density} function of the source output process $R_1(D)$ can be computed for both discrete and continuous alphabet sources from source output realizations using Arimoto and Blahut's algorithm Also, $H(X)$ {h(X)} can be estimated using histogram techniques However, a direct computation of the entropy rate {differential entropy rate} based on an estimate of the joint probability {joint probability density} function is prohibitively expensive on the data length requirement We will give algorithms for estimating order - q entropy rates and differential entropy rates In this ordering of rates, H corresponds to the first order entropy rate and h is the first

order differential entropy rate respectively. We will be concerned with the estimation of lower bounds H_2 of H and h_2 of h from source output realizations which will then be used in the estimation of $R^L(D)$

The organization of the chapter is as follows: In section 6 1, we give definitions and relevant lower and upper bounds of the rate distortion function for sources with memory. Both discrete and continuous amplitude sources are considered An algorithm to compute the first order rate distortion function, $R_1(D)$, from source output realizations is given in section 6 2 Section 6 3 is concerned with the definitions of the order - q entropy rates and differential entropy rates and their estimation procedures using the generalized correlation sum Special attention is given to the estimation of the second order entropy rate H_2 and the second order differential entropy rate h_2. In section 6 4, we give examples of the estimation of H_2 and h_2 from random process realizations. Section 6 5 gives results of the computation of the lower bound $\tilde{R}^L(D) = R_1(D) + H_2 - H(X)$ of the rate distortion function $R(D)$ for quantized speech sources

6.1 The Rate Distortion Function: Definitions and Bounds

Let us begin with a characterization of the source–user pair and a description of the notations used Consider an information 'source' S and a 'user' U For our purposes we give their characterization by the following

(a) a source output alphabet set $\mathcal{X}$ and a reproduction alphabet set $\mathcal{Y}$ We will consider two types of sources (i) a discrete alphabet source for which $\mathcal{X}$ consists of a countable number of elements and (ii) a continuous alphabet source for which $\mathcal{X}$ is the real line

(b) a 'distortion function' d $\mathcal{X} \times \mathcal{Y} \to R^+$ from the source alphabet – reproduction alphabet pairs into the set of nonnegative real numbers

(c) a probability law governing the source output It is assumed that the source output is an infinite sequence of random variables $\{X_i\}$, $-\infty < i < \infty, X_i \in \mathcal{X}$ which occur at the rate of f_s per second We assume the source output is stationary

and ergodic and is governed by a set of consistent probability {probability density} functions $p(x_1, x_2, \quad , x_n)$, $1 \leq n < \infty$, $x_i \in \mathcal{X}$ Although we have used the same notation for the joint probability function of a discrete alphabet source and the joint probability density function of a continuous alphabet source, the usage should be clear from the context For continuous alphabet sources we restrict to the case of absolutely continuous joint probability distributions

A sequence of random variables $X_i, X_{i+1}, \quad , X_{i+d-1}$ will be represented as a random vector $\mathbf{X}_i^d$

$$\mathbf{X}_i^d = [X_i \, X_{i+1} \quad X_{i+d-1}]^T \tag{6.1}$$

Likewise a sample realization of $\mathbf{X}_i^d$ will be represented by $\mathbf{x}_i^d$

$$\mathbf{x}_i^d = [x_i \, x_{i+1} \quad x_{i+d-1}]^T \tag{6.2}$$

Similarly, a d- dimensional reproduced random vector will be denoted by $\mathbf{Y}_i^d$ and its sample realization by $\mathbf{y}_i^d$

It must be remembered that signal compression is a deterministic process Its function is to replace a sequence of source output symbols with a sequence of symbols from the reproducing alphabet in such a way that the new sequence has less entropy but at the cost of some distortion The same block of source symbol sequence will always produce the same block of reproducing symbol sequence But when one restricts attention to a single source output symbol (or a relatively small subblock) without knowledge of the previous or subsequent source output symbols or of that symbol's position within a block, then the reproducing symbol y_k into which a source output symbol x_j is encoded can be thought of as a random variable even though at the block level the encoding is deterministic

For a source output sequence of length n, the mutual information (defined in eq (6.10)) between the source sequence and the reproduced sequence follows the relation

$$I(\mathbf{X}^n, \mathbf{Y}^n) = H(\mathbf{Y}^n) - H(\mathbf{Y}^n/\mathbf{X}^n) \tag{6.3}$$

where $H(\mathbf{Y}^n)$ is the entropy of the reproduced sequence of length n and $H(\mathbf{Y}^n/\mathbf{X}^n)$ is the conditional entropy of $\mathbf{Y}^n$ subject to $\mathbf{X}^n$ Since signal compression is a

deterministic process, $H(\mathbf{Y}^n/\mathbf{X}^n) = 0$ Intuitively, we want to reduce the entropy $H(\mathbf{Y}^n)$ of the reproduced symbol sequence subject to the condition on the allowable average distortion Thus, we would like to reduce the mutual information $I(\mathbf{X}^n, \mathbf{Y}^n)$ subject to this constraint.

The general rate distortion function, $R(D)$, with respect to the distortion function d_n, is defined by [12], [49]

$$R(D) = \lim_{n\to\infty} R_n(D) \tag{6 4}$$

where $R_n(D)$ is the n^{th}- order rate distortion function

$$R_n(D) = \frac{1}{n} \inf_{p(\mathbf{y}^n/\mathbf{x}^n)\in P_D} I(\mathbf{X}^n, \mathbf{Y}^n) \tag{6 5}$$

P_D is the set of conditional joint probability {joint probability density} functions $p(\mathbf{y}^n/\mathbf{x}^n)$ such that the expected distortion is $\leq D$. Thus, for discrete alphabet sources

$$P_D = \left\{ p(\mathbf{y}^n/\mathbf{x}^n) \quad \sum_{\mathbf{x}^n}\sum_{\mathbf{y}^n} p(\mathbf{x}^n)p(\mathbf{y}^n/\mathbf{x}^n)\, d_n(\mathbf{x}^n, \mathbf{y}^n) \leq D \right\} \tag{6 6a}$$

and for continuous alphabet sources,

$$P_D = \left\{ p(\mathbf{y}^n/\mathbf{x}^n) \quad \int_{\mathbf{x}^n}\int_{\mathbf{y}^n} p(\mathbf{x}^n)p(\mathbf{y}^n/\mathbf{x}^n)\, d_n(\mathbf{x}^n, \mathbf{y}^n)\mathbf{dy}^n\mathbf{dx}^n \leq D \right\} \tag{6 6b}$$

where $d_n(\mathbf{x}^n, \mathbf{y}^n)$ is a distortion criterion between sequences $\mathbf{x}^n$ and $\mathbf{y}^n$ It is called a *single letter distortion criterion* if

$$d_n(\mathbf{x}^n, \mathbf{y}^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, y_i) \tag{6 7}$$

For example, the Hamming distortion is given by

$$d(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases} \tag{6 8}$$

and the squared error distortion is given by

$$d(x, y) = (x - y)^2 \tag{6 9}$$

For discrete alphabet source–user pairs

$$I(\mathbf{X}^n, \mathbf{Y}^n) = \sum_{\mathbf{x}^n} \sum_{\mathbf{y}^n} p(\mathbf{x}^n) p(\mathbf{y}^n/\mathbf{x}^n) \log \frac{p(\mathbf{y}^n/\mathbf{x}^n)}{\sum_{\mathbf{x}^n} p(\mathbf{x}^n) p(\mathbf{y}^n/\mathbf{x}^n)} \tag{6 10a}$$

and for continuous alphabet source-user pairs

$$I(\mathbf{X}^n, \mathbf{Y}^n) = \int_{\mathbf{x}^n} \int_{\mathbf{y}^n} p(\mathbf{x}^n) p(\mathbf{y}^n/\mathbf{x}^n) \log \frac{p(\mathbf{y}^n/\mathbf{x}^n)}{\int_{\mathbf{x}^n} p(\mathbf{x}^n) p(\mathbf{y}^n/\mathbf{x}^n) d\mathbf{x}^n} d\mathbf{y}^n d\mathbf{x}^n \tag{6 10b}$$

$R(D)$ is a continuous, monotonic decreasing, convex function in the interval from $D = 0$ to $D = D_{max}$ where D_{max} is given by $R(D) = 0$ for $D \geq D_{max}$ For each $D \in (0, D_{max})$ there is one and only one relative minimum of $I(\mathbf{X}^n, \mathbf{Y}^n)$ in P_D This minimum value is $R(D)$ and it always occurs at a point having an average distortion *equal* to D The inequality part of eqs (6 6a) and (6 6b) are operative only for $D \geq D_{max}$

The analytical computation of $R(D)$ is largely an intractable problem Therefore, one resorts to computing its bounds An obvious upper bound which is usually computed is $R_1(D)$ corresponding to the marginal probability {probability density} function of the source alphabet For continuous alphabet sources, a tighter upper bound to $R(D)$ based on the mean square error criterion is provided by the rate distortion function of a Gaussian source having the same first and second order moments [12]

We are here concerned with the computation of a lower bound $R^L(D)$ of $R(D)$ [12], [151] For a discrete alphabet source,

$$R^L(D) = R_1(D) + H - H(X) \tag{6 11}$$

where $H(X)$ is the entropy of the source based on the marginal probability

$$H(X) = -\sum_x p(x) \log p(x) \tag{6 12}$$

and H is the entropy rate

$$H = \lim_{n \to \infty} \frac{1}{n} H(\mathbf{X}^n) \tag{6 13a}$$

$$= -\lim_{n \to \infty} \frac{1}{n} \sum_{\mathbf{x}^n} p(\mathbf{x}^n) \log p(\mathbf{x}^n) \tag{6 13b}$$

Similarly for a continuous alphabet source

$$R^L(D) = R_1(D) + h - h(X) \tag{6.14}$$

where $h(X)$ is the differential entropy of the source based on the marginal density function:

$$h(X) = -\int_x p(x) \log p(x)\, dx \tag{6.15}$$

and h is the differential entropy rate

$$h = \lim_{n\to\infty} \frac{1}{n} h(\mathbf{X}^n) \tag{6.16a}$$

$$= -\lim_{n\to\infty} \frac{1}{n} \int_{\mathbf{x}^n} p(\mathbf{x}^n) \log p(\mathbf{x}^n)\, \mathbf{dx}^n \tag{6.16b}$$

Successively tighter lower bounds of $R(D)$ are obtained from $R_n(D)$, the n^{th}- order rate distortion function for increasing values of n. However, we restrict attention to the computation of $R^L(D)$, eqs. (6.11), (6.14). The first order rate distortion function can be computed using Arimoto and Blahut's algorithm [16], [17]. The entropy {differential entropy} can be computed by first estimating the probability {probability density} function using standard histogram technique. Therefore we restrict attention to the estimation of entropy rate H for discrete alphabet sources and the differential entropy rate h for continuous alphabet sources. Since we estimate the second order entropy rate H_2 and the second order differential entropy rate h_2 which are lower bounds to H and h respectively, we actually compute the following lower bound to $R(D)$:

$$\tilde{R}^L(D) = R_1(D) + H_2 - H(X) \tag{6.17a}$$

for discrete alphabet sources and

$$\tilde{R}^L(D) = R_1(D) + h_2 - h(X) \tag{6.17b}$$

for continuous alphabet sources.

In the following section, we review Arimoto and Blahut's algorithm for the numerical computation of the first order rate distortion function from source output realizations.

6.2 An Algorithm for the Computation of the First Order Rate Distortion Function

In this review of Arimoto and Blahut's algorithm for the computation of $R_1(D)$ from source output realizations, we follow [16], [17] It is possible to use this algorithm to compute higher order rate distortion functions but the computational effort and data length requirement increase exponentially with the order We will restrict the following discussion to the computation of $R_n(D)$ for $n = 1$ for discrete alphabet sources only It can be appropriately extended for continuous alphabet sources also Since $R_1(D)$ makes use of the marginal probability function only, we will do away with the notation of the explicit dependence of the various functions on n For example, $p(\mathbf{x}^n)$ will be replaced by $p(x)$, $p(\mathbf{y}^n/\mathbf{x}^n)$ by $p(y/x)$, $R_1(D)$ by $R(D)$ etc compared to the definitions in section 6 1 The algorithm follows from a set of theorems due to Blahut [16] which we present, without proof, as Facts below

Fact 6.1: $R(D)$ is a decreasing, convex, and hence continuous function defined in the interval $0 \leq D \leq D_{max}$, where

$$D_{max} = \min_{p(y)} \sum_x p(x)d(x, y) \tag{6 18}$$

In this interval, the inequality constraint on D is satisfied with equality ■

D_{max} corresponds to the smallest average distortion that can be obtained with zero information transmission This is the maximum distortion that needs to be tolerated by a source Fact 6 1 tells us that the inequality constraint in the definition of $R(D)$ can be replaced with an equality constraint in $0 \leq D \leq D_{max}$ This allows us to accommodate the constraint with a Lagrange multiplier s and express $R(D)$ as

$$R(D_s) = \min_{p(y/x)} \left[\sum_x \sum_y p(x)p(y/x) \log \frac{p(y/x)}{\sum_x p(x)p(y/x)} - s \left(\sum_x \sum_y p(x)p(y/x)d(x, y) - D \right) \right] \tag{6 19}$$

where now $p(y/x)$ is unconstrained except that it must be a valid conditional probability function For each choice of the Lagrange multiplier s, the minimum will be

achieved by some conditional probability function $p^*(y/x)$ In this way, the Lagrange multiplier s becomes the independent parameter Next, the minimization problem is enlarged into a double minimization problem to facilitate the development of the algorithm

Fact 6.2: The rate distortion function $R(D)$ can be expressed as a double minimum

$$R(D) = sD + \min_{p(y)} \min_{p(y/x)} \left[\sum_x \sum_y p(x)p(y/x) \log \frac{p(y/x)}{p(y)} - s \sum_x \sum_y p(x)p(y/x)d(x,y) \right] \tag{6 20}$$

where

$$D = \sum_x \sum_y p(x)p^*(y/x)d(x,y) \tag{6 21}$$

and $p^*(y/x)$ achieves the minimum For fixed $p(y/x)$, the right side is minimized by

$$p(y) = \sum_x p(x)p(y/x) \tag{6 22}$$

For fixed $p(y)$, the right side is minimized by

$$p(y/x) = \frac{p(y) \exp(sd(x,y))}{\sum_y p(y) \exp(sd(x,y))} \tag{6 23}$$

■

The simultaneous conditions form the basis of the algorithm given in Fact 6 4

Fact 6.3: The rate distortion function can be expressed in the form

$$R(D) = \max_{s \in [-\infty, 0]} \min_{p(y)} \left[sD - \sum_x p(x) \log \sum_y p(y) \exp(sd(x,y)) \right] \tag{6 24}$$

■

This follows as a corollary from Fact 6 2 By varying parameter s from $-\infty$ to 0, one can get the entire $R(D)$ curve for $0 \leq D \leq D_{max}$

Fact 6.4: Let the parameter $s < 0$ be given and let $A(x,y) = \exp(sd(x,y))$ Let $p^0(y)$ be any initial probability function such that all components are nonzero Also, let $p^{r+1}(y)$ be given in terms of $p^r(y)$ by

$$p^{r+1}(y) = p^r(y) \sum_x \frac{p(x)A(x,y)}{\sum_y p^r(y)A(x,y)} \tag{6 25}$$

Then,

$$D(p^r(y/x)) \longrightarrow D_s, \quad r \to \infty \tag{6 26}$$

and

$$I(p(x), p^r(y/x)) \longrightarrow R(D_s) \quad \text{as} \quad r \to \infty \tag{6 27}$$

where

$$p^r(y/x) = \frac{p^r(y)A(x,y)}{\sum_y p^r(y)A(x,y)} \tag{6 28}$$

and $(D_s, R(D_s))$ is a point on the $R(D)$ curve parametrized by s ■

In eq (6 26) we have denoted the explicit dependence of D at the r^{th} iteration on $p^r(y/x)$ In eq (6 27), $I(X,Y)$ has been denoted by $I(p(x), p^r(y/x))$ to highlight the dependence on $p(x)$ and $p^r(y/x)$ The application of Fact 6 4 to the computation of the rate distortion function is illustrated in fig 6 1 The termination of the algorithm is based on Fact 6 5 given below It gives upper and lower bounds on $R(D)$ and is helpful in estimating the residual error after each iteration

Fact 6.5: Let the parameter $s < 0$ be given, and let $A(x,y) = \exp(sd(x,y))$ Suppose $p(y)$ is any reproduced alphabet probability function, and let

$$c(y) = \sum_x p(x) \frac{A(x,y)}{\sum_y p(y)A(x,y)} \tag{6 29}$$

Then, at the point

$$D = \sum_x \sum_y p(x) \frac{p(y)A(x,y)}{\sum_y p(y)A(x,y)} d(x,y) \tag{6 30}$$

we have

$$R(D) \leq sD - \sum_x p(x) \log \sum_y p(y)A(x,y) - \sum_y p(y)c(k) \log c(k) \tag{6 31}$$

and

$$R(D) \geq sD - \sum_x p(x) \log \sum_y p(y)A(x,y) - \max_y \log c(y) \tag{6 32}$$

■

This completes the development of the algorithm steps We have used this algorithm to compute $R_1(D)$ which is required in the expression of $\tilde{R}^L(D)$, eq (6 17)

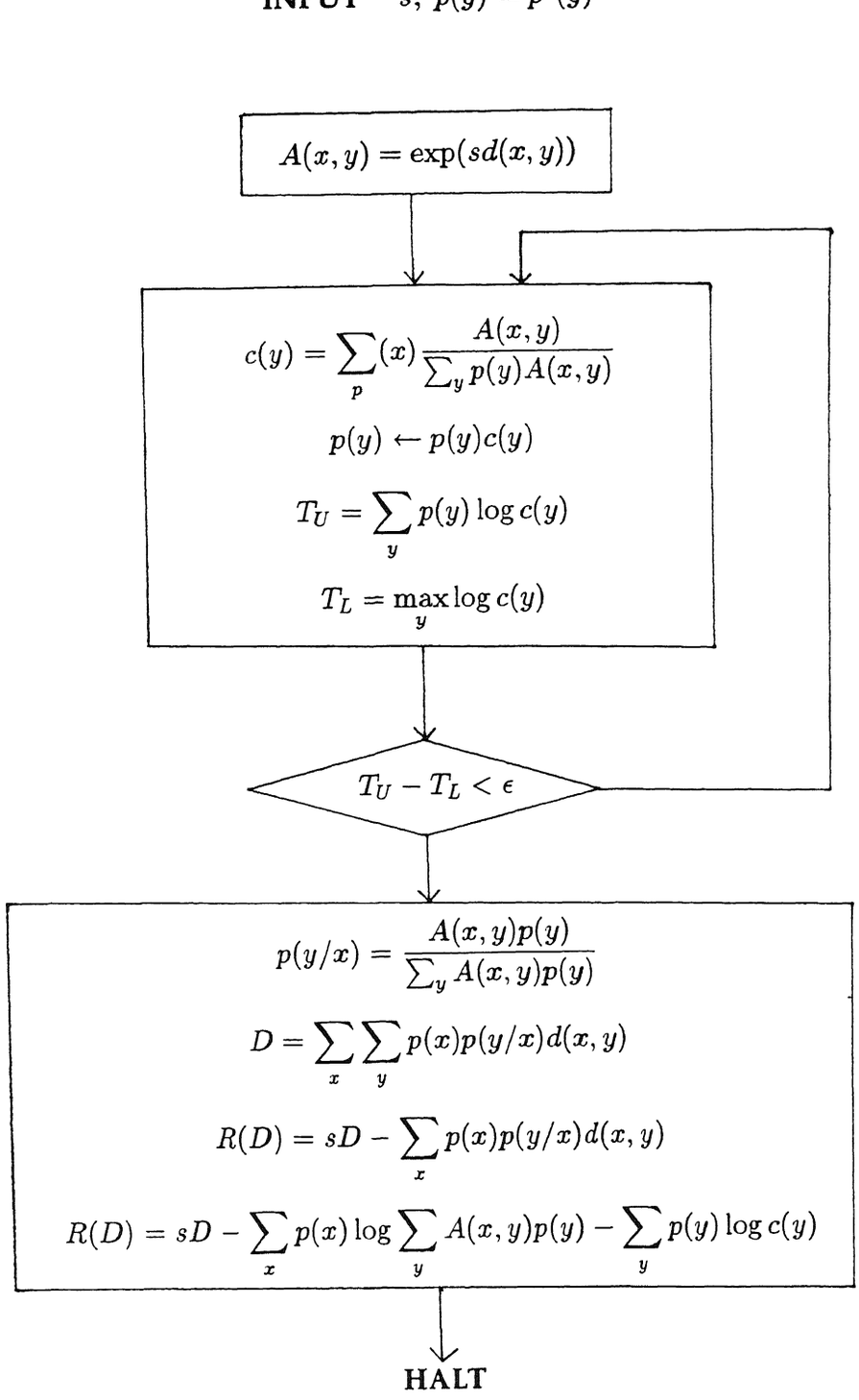


Fig. 6.1: Arimoto and Blahut's algorithm for the computation of first order rate distortion function

6.3 Computation of Lower Bounds of the Entropy Rate and Differential Entropy Rate using the Correlation Sum

A direct method for the numerical estimation of entropy and differential entropy rates is to first estimate the joint probability or probability density functions used in their expression. However, a major disadvantage of this method is the requirement of exponentially increasing data length N as higher order joint probability {probability density} functions are considered. One way of reducing the effect of this limitation is to estimate the entropy and differential entropy rates from the generalized correlation sum. The method of correlation sum has been studied and used in the estimation of dimension and metric entropy of nonlinear dynamical systems (see eg. [50], [112], [57], [58], [110], [31], also chapter 3). We will use an analogous scheme to estimate a lower bound H_2 of the entropy rate and extend it to estimate a lower bound h_2 of the differential entropy rate.

We first consider the estimation of order-q entropy rates of a *continuous* alphabet stationary ergodic source with memory in the following subsection. Thereafter, we discuss the estimation of the second order entropy rate H_2 for a discrete alphabet source and the second order differential entropy rate h_2 for a continuous alphabet source in succeeding subsections. Ideally, H [illegible] for a continuous alphabet source. If the divergence of the entropy rate is avoided by considering a finite partition, the result is a reflection of the choice of partition resolution. The necessity for considering a partition resolution r away from 0 is due to the finite resolution and limited length of the time-series data available in a numerical or experimental situation.

6.3.1 Generalized Entropy Rates and their Estimation using the Generalized Correlation Sum

Consider a discrete time information source S whose output sequence is governed by a probability law as given in section 6.1. Let $\mathcal{X}$ be the real line or a continuous subset of it. To define the entropy rate H, partition the support $\mathcal{X}^n$ of $\mathbf{X}^n$ into $\Pi_{(r,n)} = [\pi_1, \pi_2, \ldots, \pi_{M(r,n)}]$ of side resolution r. Here $M(r,n)$ denotes the number of

partitions of the support of $\mathbf{X}^n$ Let P_{π_l} be the normalized probability measure of the partition π_l Then

$$H(\mathbf{X}^n, r) = -\sum_{l=1}^{M(r,n)} P_{\pi_l} \log P_{\pi_l}, \quad n = 1, 2, \tag{6 33}$$

is the entropy based on the n-dimensional joint density function at resolution r,

$$H(r) = \lim_{n\to\infty} \frac{1}{n} H(\mathbf{X}^n, r) \tag{6 34}$$

is the entropy rate at partition resolution r and

$$H = \lim_{r\to 0} H(r) \tag{6 35}$$

is the entropy rate of source S

In order to consider the estimation of a lower bound of H, we consider the more general "order-q Renyi entropy rate" or "generalized entropy rate", H_q [112], [116] The order-q entropy of dimension n at resolution r is given by

$$H_q(\mathbf{X}^n, r) = -\frac{1}{q-1} \log \sum_{l=1}^{M(r,n)} P_{\pi_l}^q, \quad q \neq 1 \tag{6 36a}$$

$$H_1(\mathbf{X}^n, r) = H(\mathbf{X}^n, r) = -\sum_{l=1}^{M(r,n)} P_{\pi_l} \log P_{\pi_l} \tag{6 36b}$$

Then,

$$H_q(r) = \lim_{n\to\infty} \frac{1}{n} H_q(\mathbf{X}^n, r) \tag{6 37}$$

is the *generalized entropy rate at resolution* r and

$$H_q = \lim_{r\to 0} H_q(r) \tag{6 38}$$

is the *generalized entropy rate*

Let us now consider the estimation of the generalized entropy rate from the generalized correlation sum There is an important approximation that relates the generalized correlation sum to $\sum_{l=1}^{M(r,n)} P_{\pi_l}^q$ in eq (6 36) above Let us consider this approximation and the subsequent relation of H_q to the generalized correlation sum

Let $\mathcal{B}_{\mathbf{x}^n}(r,n)$ represent a ball of radius r around the point $\mathbf{x}^n$ in $\mathcal{X}^n$, the support of $\mathbf{X}^n$. Also, let its probability be given by

$$B_{\mathbf{x}^n}(r,n) = P_{\mathcal{B}_{\mathbf{x}^n}(r,n)} \tag{6.39}$$

Given a sample realization $x_1, x_2, \ldots, x_{N'}$ of the source output, construct vectors

$$\mathbf{x}_i^n = [x_i \; x_{i+1} \; \ldots \; x_{i+n-1}]^T, \quad i = 1, 2, \ldots, N' - n + 1 \tag{6.40}$$

For a point $\mathbf{x}_j^n$, an estimate of $B_{\mathbf{x}_j^n}(r,n)$ may be obtained from

$$\hat{B}_{\mathbf{x}_j^n} = \frac{1}{N-1} \sum_{\substack{i=1 \\ i \neq j}}^{N} \Theta\left(r - \|\mathbf{x}_i^n - \mathbf{x}_j^n\|\right) \tag{6.41}$$

where $N = N' - n + 1$, $\mathbf{x}_i^n$ is given by eq. (6.40) and $\Theta(\arg) = 1$ for $\arg \geq 0$ and $\Theta(\arg) = 0$ otherwise. The choice of the metric determines, for example, whether $\mathcal{B}_{\mathbf{x}^n}(r,n)$ is a n-dimensional sphere (L_2 norm) or a cube (L_∞ norm) etc. However, H_q itself is invariant to the choice of the metric.

Now let $P_{\pi_{l(j)}}$ be the estimation of the probability of that partition π_l which contains the j^{th} point $\mathbf{x}_j^n$. If this partition contains $N_{\pi_{l(j)}}$ points, then the important approximation is

$$\hat{B}_{\mathbf{x}_j^n} \simeq \frac{N_{\pi_{l(j)}}}{N} = P_{\pi_{l(j)}} \tag{6.42}$$

The idea behind this approximation is that most points in $\pi_{l(j)}$ will be within r of $\mathbf{x}_j^n$ and although some points that are further away will be ignored, others that are close enough but in neighbouring partitions will be counted. The error in approximation is only a factor of order unity [50].

The estimated generalized correlation sum is given by

$$\hat{C}_q(r,n,N) = \frac{1}{N} \sum_{j=1}^{N} \left[\hat{B}_{\mathbf{x}_j^n}(r,n)\right]^{q-1}, \quad q \neq 1 \tag{6.43a}$$

$$\hat{C}_1(r,n,N) = \lim_{q \to 1^+} \hat{C}_q(r,n,N) \tag{6.43b}$$

The following derivation shows the approximation between the generalized correlation sum and the probability of partitions. It proceeds by substituting $P_{\pi_{l(j)}}$ for

$\hat{B}_{\mathbf{x}_i^n}(r,n)$ and replacing the sum over points $i = 1, 2, \ldots, N$ with a sum over the partitions and a sum over the $\mathbf{x}_j^n$, $j = 1, 2, \ldots, N$ in each partition

$$\begin{aligned}\hat{C}_q(r,n,N) &= \frac{1}{N}\sum_{j=1}^{N}\left[\hat{B}_{\mathbf{x}_j^n}(r,n)\right]^{q-1} \\ &\simeq \frac{1}{N}\sum_{j=1}^{N}\left[P_{\pi_{l(j)}}\right]^{q-1} \\ &= \frac{1}{N}\sum_{l=1}^{M(r,n)}\sum_{j=1}^{N} I_{\pi_l}(\mathbf{x}_j^n)\,[P_{\pi_l}]^{q-1} \\ &= \frac{1}{N}\sum_{l=1}^{M(r,n)} N_{\pi_l}\,[P_{\pi_l}]^{q-1} \\ &= \sum_{l=1}^{M(r,n)} [P_{\pi_l}]^q \end{aligned} \tag{6 44}$$

Here $I_S(\mathrm{s}) = 1$ if $\mathrm{s} \in S$ and 0 otherwise Comparing eqs (6 36)–(6 38) and eq (6 44) we have an estimate of the generalized entropy rate

$$\hat{H}_q(\mathbf{X}^n, r) = -\frac{1}{q-1}\log \hat{C}_q(r,n,N), \quad q \neq 1 \tag{6 45a}$$

$$\hat{H}_1(\mathbf{X}^n, r) = \hat{C}_1(r,n,N) \tag{6 45b}$$

$$\hat{H}_q(r) = \lim_{n\to\infty} \frac{1}{n}\hat{H}_q(\mathbf{X}^n, r) \tag{6 46}$$

$$\hat{H}_q = \lim_{r\to 0} \hat{H}_q(r) \tag{6 47}$$

Note that we have not mentioned the explicit dependence of the estimates on the data length N Alternatively,

$$\hat{H}_q = \lim_{r\to 0}\lim_{n\to\infty}\left[\frac{1}{q-1}\log\frac{\hat{C}_q(r,n,N)}{\hat{C}_q(r,n+1,N)}\right], \quad q \neq 1 \tag{6 48a}$$

$$\hat{H}_1 = \lim_{q\to 1^+} H_q \tag{6 48b}$$

The above equation can be used to estimate the entropy rate H by approaching the limit $q \to 1^+$ However, we will estimate a particularly easy and special case of the order-q entropy rate, i e , for $q = 2$ The second order entropy rate H_2 is a lower

bound of the entropy rate H, i e , $H_2 \leq H$ For this case, the correlation sum is given by

$$\hat{C}(r,n,N) = \hat{C}_2(r,n,N)$$
$$= \frac{1}{N(N-1)} \sum_{j=1}^{N} \sum_{\substack{i=1 \\ i \neq j}}^{N} \Theta\left(r - \|\mathbf{x}_i^n - \mathbf{x}_j^n\|\right) \tag{6 49}$$

where $\mathbf{x}_i^n, \mathbf{x}_j^n$ are given by eq (6 40) and $\hat{C}(r,n,N)$ is the estimated correlation sum Also,

$$\hat{H}_2 = \lim_{r \to 0} \lim_{n \to \infty} \log \frac{\hat{C}(r,n,N)}{\hat{C}(r,n+1,N)} \tag{6 50}$$

Some remarks on the above procedure are as follows

(1) Conjectured advantages of the correlation sum technique over the histogram technique in estimating the entropy rate: The correlation sum $\hat{C}(r,n,N)$ computes the fraction of the total number of pair of points of the form $\mathbf{x}_i^n$, $i = 1,2, \quad ,N$ for specified dimension n, that are less than or equal to a distance r of each other The conjectured advantages of this estimation procedure are

(a) Using the correlation sum technique one has access to N^2 pair of points as opposed to N points in the histogram technique Thus, in the former one can probe down to distance scales of $O(N^{-2/n})$ i e., upto the smallest interpoint distance whereas in the latter one can probe only upto $O(N^{-1/n})$ for data length N and vector dimension n This is advantageous since the entropy rate has to be evaluated at small partition resolution r.

(b) Consider the estimation procedure using both schemes at a fixed resolution r Using the histogram technique, estimates of P_{π_i} for the partition π_i that are outlying can be grossly inaccurate This is due to the finiteness of data length and can lead to an underestimate of H If we use the correlation sum technique, then the sum over the partitions is changed to a sum over points on the vector sequence $\mathbf{x}_i^n$, $i = 1,2, \quad ,N$ We find the fraction of the total number of pair of points which are within a distance r of *each other* Therefore, intuitively the inaccuracy due to finiteness of data will be less in this case

(2) Statistical properties of the correlation sum estimator: Throughout this work we are concerned with *functions* of the estimated correlation sum Therefore, it is worthwhile to know the statistical properties of the correlation sum estimator $\hat{C}(r, n, N)$ given by eq. (6 49) It estimates

$$C(r, n) = E\left[B_{\mathbf{x}^n}(r, n)\right] \tag{6 51}$$

where $B_{\mathbf{x}^n}(r, n)$ is given by eq (6 39) More specifically, $C(r, n)$ is the correlation function

$$C(r, n) = \int_{\|\Delta\| \leq r} c(\mathbf{\Delta}, n) \mathbf{d}\mathbf{\Delta} \tag{6 52a}$$

$$c(\mathbf{\Delta}, n) = \int_{\mathbf{x}_1^n} \int_{\mathbf{x}_2^n} d\mu(\mathbf{x}_1^n) d\mu(\mathbf{x}_2^n) \delta^n(\mathbf{x}_1^n - \mathbf{x}_2^n - \Delta) \tag{6 52b}$$

$c(\mathbf{\Delta}, n)$ is the probability density for the difference between two random vectors $\mathbf{x}_1^n$ and $\mathbf{x}_2^n$ both chosen according to their probability measure.

The statistical properties of the correlation sum estimator have been investigated in the context of the estimation of dimensions and metric entropy in nonlinear dynamics. It is shown in [55] that if the sample data points $\{x_1, x_2, \quad , x_N\}$ are independent, then the correlation sum estimator is unbiased. That is

$$E\left[\hat{C}(r, n, N)\right] = C(r, n) \tag{6 53}$$

Otherwise, if the data is autocorrelated, then the estimator is asymptotically unbiased However, the finite N bias can be corrected by a slight modification of the definition of the estimator [134]

$$\hat{C}(r, n, N) = \frac{1}{N(N - W)} \sum_{j=1}^{N} \sum_{\substack{i=1 \\ |i-j|>W}}^{N} \Theta\left(r - \|\mathbf{x}_i^n - \|\mathbf{x}_j^n\|\right) \tag{6 54}$$

where W is a small positive integer value Theiler has shown that the correlation sum estimator is a consistent estimator of the correlation function [138] The standard deviation σ_C of the estimator generically scales as $O(1/\sqrt{N})$ for $N \to \infty$ Those sources for which the pointwise mass function, eq (6 39), remains constant, i e ,

$$B_{\text{rms}}^2 = E[B - E(B)]^2 = 0 \tag{6 55}$$

σ_C scales as $O(1/N)$ for $N \to \infty$

(3) Implementation Aspects and Choice of Resolution Scale: Given a source output realization x_i, $i = 1, 2, \quad , N'$ the implementation consists of the following steps

(a) Construct sample vectors $\mathbf{x}_i^n$, $i = 1, 2, \quad , N$, $(N = N' - n + 1)$ and for $n = 1, 2,$ using eq (6 40)

(b) At fixed resolution r, obtain $\hat{C}(r, n, N)$ for increasing values of n using eq (6 49)

(c) Plot $\hat{H}_2(r, n)$ vs n where

$$\hat{H}_2(r, n) = \log \frac{\hat{C}(r, n, N)}{\hat{C}(r, n+1, N)} \tag{6 56}$$

The quantity $\hat{H}_2(r, n)$ should converge with increasing values of n which is read off as the estimated value $\hat{H}_2(r)$ of the second order entropy rate $H_2(r)$

Ideally, $\hat{H}_2(r)$ is evaluated at $r \to 0$ limit However, there are limitations in approaching this limit A lower limit to r is set by the accuracy of the data At distance scales of the order of the smallest interpoint distance, the correlation sum scales as a collection of isolated points Further, at small scales, the correlation sum may be inaccurate due to the presence of additive noise On the other hand, at large scales the correlation sum tends to saturate Thus, the distance scale r should be chosen between these extremes. Preferably, $\hat{H}_2(r)$ should be computed for different values of r in the intermediate range In the examples of section 6 4 it is seen that the variation of the estimate is small over a large range of values of r For more details on the implementation aspects of the correlation sum algorithm, the reader can see section 3 7, where the discussion is in the context of the estimation of correlation dimension and second order metric entropy from time series

6.3.2 Second Order Entropy Rate for a Discrete Alphabet Source

Consider the discrete alphabet source S described in section 6 1 Let $\mathcal{X} = \{a_1, a_2, \quad , a_J\}, |\mathcal{X}| = J$. Let us now represent the source alphabet set by an integer valued

set $Z = \{i \quad 1 \leq i \leq |\mathcal{X}|\}$ such that a source output sequence $\{x_i, \ -\infty < i < \infty\}$ is represented by $\{z_i, \ -\infty < i < \infty\}$ where $z_i = k$ if $x_i = a_k$, $1 \leq k \leq J$ Thus,

$$p(x_1 = a_i, x_2 = a_j, \quad , x_n = a_k) = p(z_1 = i, z_2 = j, \quad , z_n = k), \quad 1 \leq n < \infty, \ x_i \in \mathcal{X} \tag{6 57}$$

This transformation of the alphabet set does not change the generalized entropy rate because it is coordinate invariant. The second order entropy rate is estimated using

$$\hat{H}_2(r) = \lim_{n \to \infty} \log \frac{\hat{C}(r, n, N)}{\hat{C}(r, n+1, N)} \tag{6 58}$$

$$\hat{H}_2 = \lim_{r \to 1} \hat{H}_2(r) \tag{6 59}$$

Here $\hat{C}(r, n, N)$ is given by eq (6 49) In section 6 4, we give results of the computation of $\hat{H}_2$ for two specific discrete alphabet stochastic processes

6.3.3 Second Order Differential Entropy Rate for Continuous Alphabet Source

Analogous to the development in section 6 3 1, we now consider the estimation of a lower bound h_2 of the differential entropy rate h, given by eq (6 16) We begin with what we call the "order-q differential entropy rate" h_q Consider a continuous alphabet source S as described in section 6 1 The order-q differential entropy for a n-dimensional random vector $\mathbf{X}^n = [X_1, X_2, \quad , X_n]^T$ is given by

$$h_q(\mathbf{X}^n) = \int_{\mathbf{x}^n} [p(\mathbf{x}^n)]^q \ d\mathbf{x}^n, \quad q > 0, \quad q \neq 1 \tag{6 60a}$$

$$h_1(\mathbf{X}^n) = h(\mathbf{X}^n) = \int_{\mathbf{x}^n} p(\mathbf{x}^n) \log p(\mathbf{x}^n) d\mathbf{x}^n \tag{6 60b}$$

provided the integrals exist The order-q differential entropy rate is then defined as

$$h_q = \lim_{n \to \infty} \frac{1}{n} h_q(\mathbf{X}^n), \quad q > 0, q \neq 1 \tag{6 61a}$$

$$h_1 = h = \lim_{n \to \infty} \frac{1}{n} h(\mathbf{X}^n) \tag{6 61b}$$

Renyi has proved a limiting relation between the order-q entropy and the order-q differential entropy by considering the continuous distribution of the random vector

$\mathbf{X}^n$ to be a limiting case of the discrete distribution of an associated random vector $\mathbf{X}_r^n$ (see [116], Appendix)

Consider an n-dimensional random vector $\mathbf{X}^n = [X_1, X_2, \quad, X_n]^T$ with an absolutely continuous distribution Let $p(x_1, x_2, \quad, x_n)$ be the corresponding density function of $\mathbf{X}^n$ Let $\mathbf{X}_r^n = [X_{1r}, X_{2r}, \quad, X_{nr}]^T$ be an associated discrete random vector where $X_{ir} = r\left[\frac{X_i}{r}\right]$, $i = 1, 2, \quad, n$ (Here $[y]$ denotes the largest integer $\leq$ y) If $H(\mathbf{X}^n, r)$ is finite for $r = 1$, then

$$h_q(\mathbf{X}^n) = \lim_{r \to 0} [H_q(\mathbf{X}_r^n) + n \log r], \quad q > 0, \quad q \neq 1 \tag{6 62a}$$

where $H_q(\mathbf{X}_r^n, r)$ is given by eq (6 36a) and $h_q(\mathbf{X}^n)$ is given by eq (6 60a) and

$$h_1(\mathbf{X}^n) = \lim_{r \to 0} [H_1(\mathbf{X}_r^n, r) + n \log r] \tag{6 62b}$$

where $H_1(\mathbf{X}_r^n, r)$ is given by eq (6 36b) and $h_1(\mathbf{X}^n)$ is given by eq (6 60b)

This result can be extended to obtain a relation between the two entropy *rates*

$$h_q = \lim_{r \to 0} [H_q(r) + \log r], \quad q > 0, \quad q \neq 1 \tag{6 63}$$

where $H_q(r)$ is the order-q entropy rate of $\mathbf{X}^n$ at resolution scale r, eq (6 37) and h_q is the order-q differential entropy rate of $\mathbf{X}^n$, eq (6 61)

To obtain an estimate $\hat{h}_q$ of $h_q, q > 0$ we compute the right side of eq (6 63) $H_q(r)$ is estimated by $\hat{H}_q(r)$ using the generalized correlation sum at a distance scale r Thus,

$$\hat{h}_q = \lim_{r \to 0} [\hat{H}_q(r) + \log r], \quad q > 0 \tag{6 64}$$

We will estimate $\hat{h}_q$ for $q = 2$ The second order differential entropy rate h_2 is a lower bound to the differential entropy rate h_1 The estimate $\hat{h}_2$ is given by

$$\begin{aligned} \hat{h}_2 &= \lim_{r \to 0} \left[\hat{H}_2(r) + \log r\right] \\ &= \lim_{r \to 0} \log \frac{r\hat{C}(r, n, N)}{\hat{C}(r, n+1, N)} \end{aligned} \tag{6 65}$$

where $\hat{C}(r, n, N)$ is given by eq (6 49)

The choice of the resolution scale r and the implementation steps discussed in section 6 3 1 carry over to the estimation of $\hat{h}_2$. Step (c) and eq. (6 56) are modified to the following

Obtain a plot of $\hat{h}_2(r,n)$ vs n where

$$\hat{h}_2(r,n) = \log \frac{r\hat{C}(r,n,N)}{\hat{C}(r,n+1,N)} \tag{6 66}$$

The quantity $\hat{h}_2(r,n)$ should converge to $\hat{h}_2(r)$ with increasing values of n and for fixed r which is then read off as the estimated value $\hat{h}_2$ of the second order differential entropy rate

6.4 Results of the Estimation of Second Order Entropy and Differential Entropy Rates from Time Series Realizations

We consider two examples each of the estimation of the second order entropy rate H_2 and the second order differential entropy rate h_2 using their respective estimates based on the correlation sum In each case the Euclidian distance is used as the metric in the estimation of the correlation sum in eq (6.49)

1. Discrete Alphabet i.i.d. Random Process: Consider an i i d random process $\{X_i\}$ whose random variables are drawn from a discrete alphabet set $\mathcal{X} = \{1,2,3,4\}$ Let $p(i) = \frac{1}{4}$, $i \in \mathcal{X}$ We have

(a) Entropy rate $H = 2\,0$ bits/sample

(b) Estimated value $\hat{H}_2(r) = 1\,88 \pm 0\,03$ bits/sample at $r^2 = 2\,0$ over 8 realizations

(c) Estimated value $\hat{H}_2 = 1\,80 \pm 0\,10$ bits/sample over 4 values of r^2 from 1 0 to 4 0 at intervals of 1 0.

2. Discrete Markov Chain: Let $\{X_i\}$ be an 8 state stationary Markov chain as shown in fig 6 2 Its probability transition matrix P has entries P_{ij} given by

$$P_{ij} = p(X_{n+1} = j/X_n = i) = \begin{cases} 1/2, & \text{if } i = j \\ 1/4, & \text{if } i \ominus_8 j = \pm 1,\ i,j \in \{0,1, \quad ,7\} \end{cases} \tag{6 67}$$

For a Markov chain

$$H = \lim_{n\to\infty} \frac{1}{n} H(\mathbf{X}^n) = \lim_{n\to\infty} H(X_n/X_{n-1}, \quad ,X_1) = H(X_2/X_1) \tag{6 68}$$

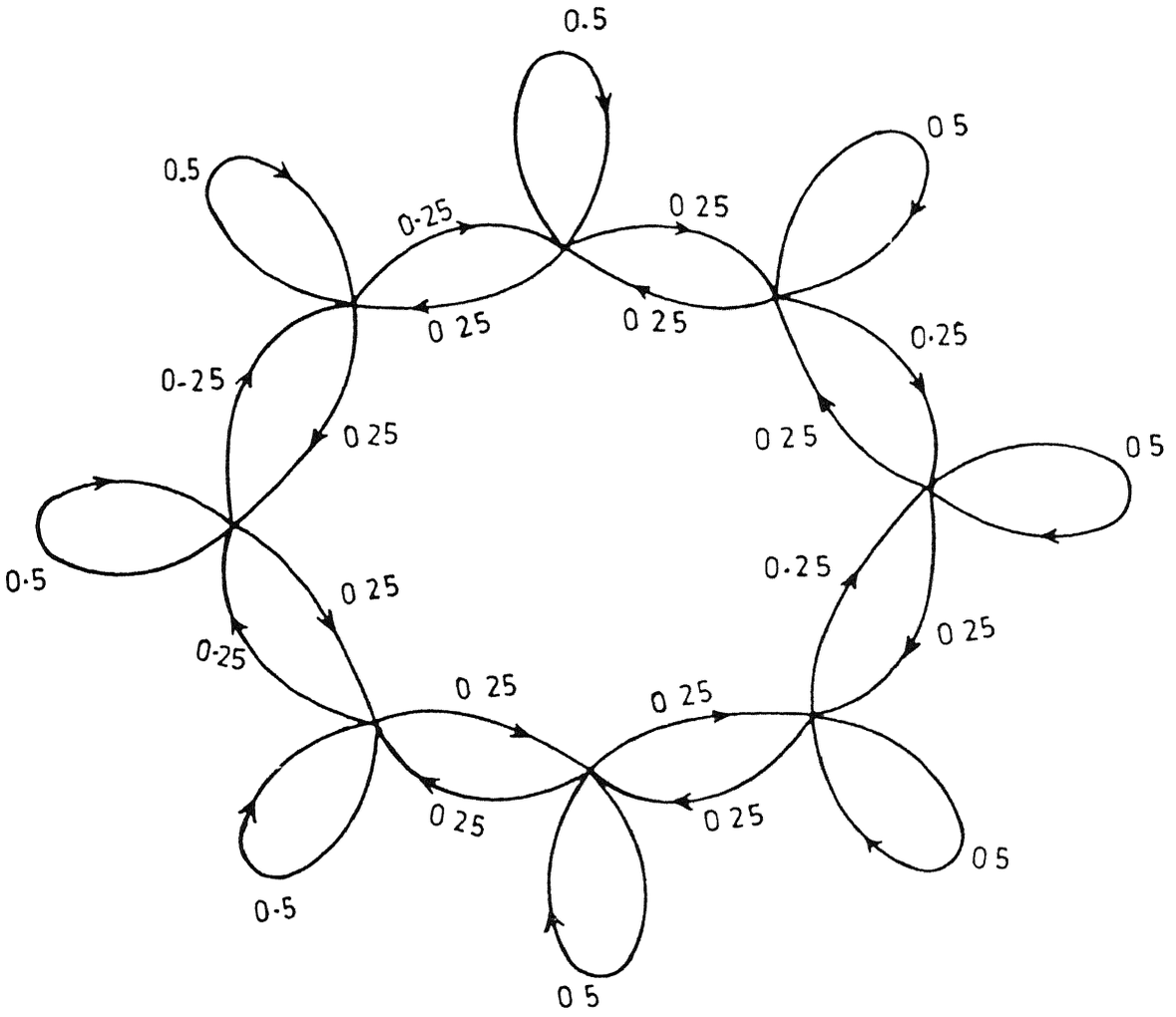


Fig. 6.2: An 8 state Markov chain for example 2, section 6.4

We have·

(a) $H = 1\,5$ bits/sample.

(b) $\hat{H}_2(r) = 1\,34 \pm 0\,03$ bits/sample at $r^2 = 2\,0$ over 8 realizations

(c) $\hat{H}_2 = 1\,31 \pm 0\,08$ bits/sample over 8 values of r^2 from 1 0 to 8 0 at intervals of 1 0

Figure 6 3 shows the convergence of $\hat{H}_2(r, n)$ with increasing dimension n for one realization each of examples 1 and 2

3. Gaussian Distributed i.i.d. Random Process: Let $\{e_n\}$ be a white Gaussian noise process where $e_n \sim \mathcal{N}(0, 1)$ For such a process the differential entropy rate is given by $h = \frac{1}{2}\log_2(2\pi e\sigma_e^2)$ (see eg [12], [32]) Thus, we have

(a) Differential entropy rate $h = 2\,047$ per sample

(b) Estimated value $\hat{h}_2(r) = 1\,85 \pm 0\,02$ per sample at $r = 0\,3\sigma_e$ over 8 realizations

(c) $\hat{h}_2 = 2\,03 \pm 0\,17$ per sample over 10 values of r from $0\,1\sigma_e$ to $1\,0\sigma_e$ at intervals of $0\,1\sigma_e$

4. Jointly Gaussian Distributed Random Process: Let X_i be a jointly Gaussian distributed random process with n-dimensional random vectors $\mathbf{X}^n = [X_1, X_2, \quad, X_n]^T$ distributed as $\mathcal{N}_n(\mathbf{m}, K_n)$ where $\mathbf{m}$ is the mean vector and K_n is the $n \times n$ covariance matrix The differential entropy rate of the process is given by

$$h = \frac{1}{2}\log(2\pi e) + \lim_{n\to\infty} \frac{1}{2n}\log|K_n| \tag{6 69}$$

where $|K_n|$ is the determinant of K_n [12], [32]

Consider a zero mean Gaussian random process with

$$E[X_i X_{i+k}] = \begin{cases} \sigma_X^2, & k = 0 \\ a\sigma_X^2, & k = \pm 1 \\ 0, & \text{otherwise} \end{cases} \tag{6 70}$$

For $\sigma_X^2 = 2\,0$ and $a = 0\,5$, we have

(a) $h = 2\,047$ per sample

(b) $\hat{h}_2(r) = 1\,87 \pm 0\,02$ per sample at $r = 0\,3\sigma_e$ over 8 realizations

(c) $\hat{h}_2 = 2\,16 \pm 0\,21$ per sample over 10 values of r from $0\,1\sigma_e$ to $1\,0\sigma_e$ at intervals of $0\,1\sigma_e$

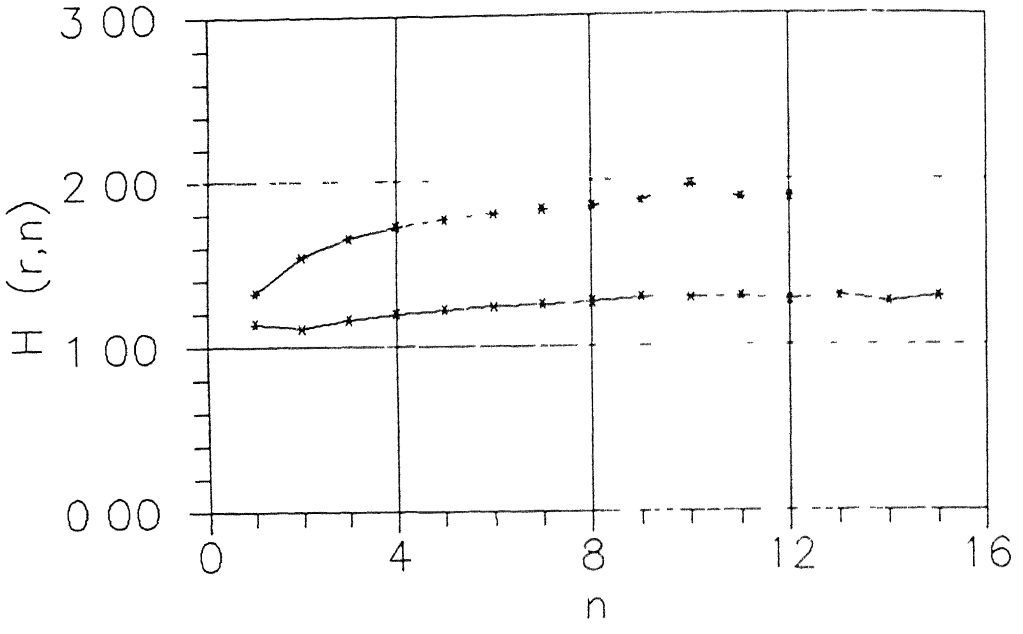


Fig. 6.3: Graph of $\hat{H}_2(r,n)$ vs n, eq (6 56) The estimation is done for one realization each of examples 1 and 2, section 6 4

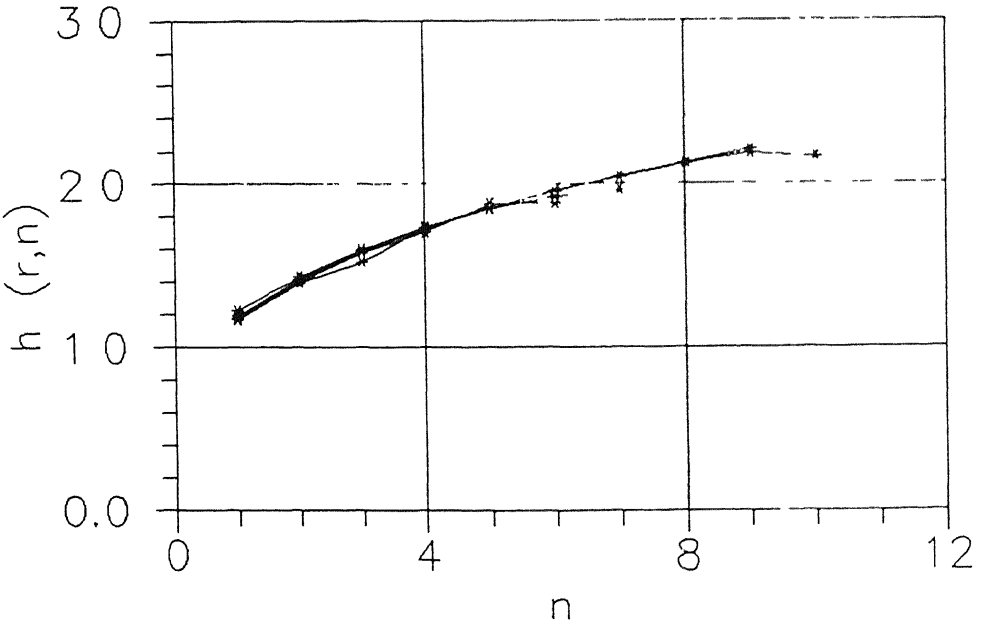


Fig. 6.4: Plots of $\hat{h}_2(r,n)$ vs n, eq (6 66) for 4 different distance scales, $r = 0\,2\sigma_e, 0\,4\sigma_e$, $0\,6\sigma_e$ and $0\,8\sigma_e$ for example 3, section 6 4 Very little variation in $\hat{h}_2$ is observed as a function of r

A data length $N = 40000$ was used for each realization in parts (b) and (c) in all the 4 examples above

It must be emphasised that we have computed a lower bound to the entropy and differential entropy rates in the above examples. Nevertheless the estimated values are within 11% of the respective theoretical values of the rate. As the examples show, the variation of $\hat{H}_2$ and $\hat{h}_2$ is small over a fairly large range of resolution scales r Figure 6.4 shows the variation of $\hat{H}_2(r,n)$ vs n for 4 different resolution scales $r = 0\,2\sigma_e$, $0.4\sigma_e$, $0\,6\sigma_e$ and $0\,8\sigma_e$ for example 3 Thus, while the choice of r may not be an important factor in the estimation procedure, we suggest that the estimate be computed for a range of resolution scales and reported as an average.

6.5 Computation of a Lower bound of R(D) for Quantized Speech Source

Various rate distortion functions for speech sources have been proposed previously These include the first order rate distortion function based on *adhoc* speech probability models such as the Gaussian and Laplacian density functions, the rate distortion function based on the assumption of a Gaussian AR source using the mean squared error distortion criterion and those based on perceptually significant distortion measures See eg [2], [122], [77]

In this section, we give results of the computation of the lower bound $\tilde{R}^L(D)$ of the rate distortion function $R(D)$ for quantized speech sources with respect to the mean squared error distortion criterion. (Note that in a strict sense the algorithm is not applicable because speech is a time varying source) We have

$$\tilde{R}^L(D) = R_1(D) + H_2 - H(X) \qquad (6\,17a)$$

where $R_1(D)$ is the first-order rate distortion function, eq (6 5), $H(X)$ is the first order entropy, eq. (6 12), based on the marginal distribution of the quantized speech source and H_2 is the second order entropy rate given by eqs (6 36)–(6 38), and estimated from time series data using eq (6 50)

The speech database is as described in Appendix B (database 2) It consists of 4 *phoneme specific* sentences spoken by 2 males and 2 females The speech signal was

prefiltered at 3 9 kHz and sampled at 8 kHz at a resolution of 16 bits/sample A total of 39 96s duration of speech data has been used for this work

The lower bound $\tilde{R}^L(D)$ of $R(D)$ has been computed for quantized speech at 3 resolutions of 6, 8 and 10 bits/sample The first order rate distortion function, $R_1(D)$, has been computed using Arimoto and Blahut's algorithm Figure 6 5 shows graphs of the first order rate distortion function for the 3 quantization levels with respect to the squared error distortion criterion The first order entropy, $H(X)$, has been computed by approximating the probability function with histograms See Table 6 1 for the results of the estimation of $H(X)$ and H_2 Figure 6 6 shows the convergence of $\hat{H}_2(r, n)$ with increasing dimension n obtained for two sentence utterances - one each by a male and female speaker Finally, $R^L(D)$ has been plotted as rate vs SNR in fig 6 7 for quantization resolutions of 6, 8 and 10 bits/sample It is seen that the graphs for different speech source resolutions intersect each other This is because two contradicting factors determine the rate A higher source resolution implies that greater information needs to be transmitted However, higher quantization resolution also gives the scope for further removing the redundancy between the source output symbols thereby reducing the rate based on the given distortion criterion

The above example illustrates the procedure for computing $\tilde{R}^L(D)$ However, to evaluate the performance of a *low* bit rate speech coder with respect to a lower bound, it is more appropriate to use a perceptually significant distortion criterion

Source Resolution (bits/sample)	Entropy, $\hat{H}(X)$ (Kbps)	Second Order Entropy Rate, $\hat{H}_2(\times 10^1)$
6	21 2	1 34
8	35 35	1 47
10	50 84	2 20

Table 6.1· (1) The estimated entropy using histogram technique, and (2) the estimated second order entropy rate, for quantized speech sources at 3 resolution scales

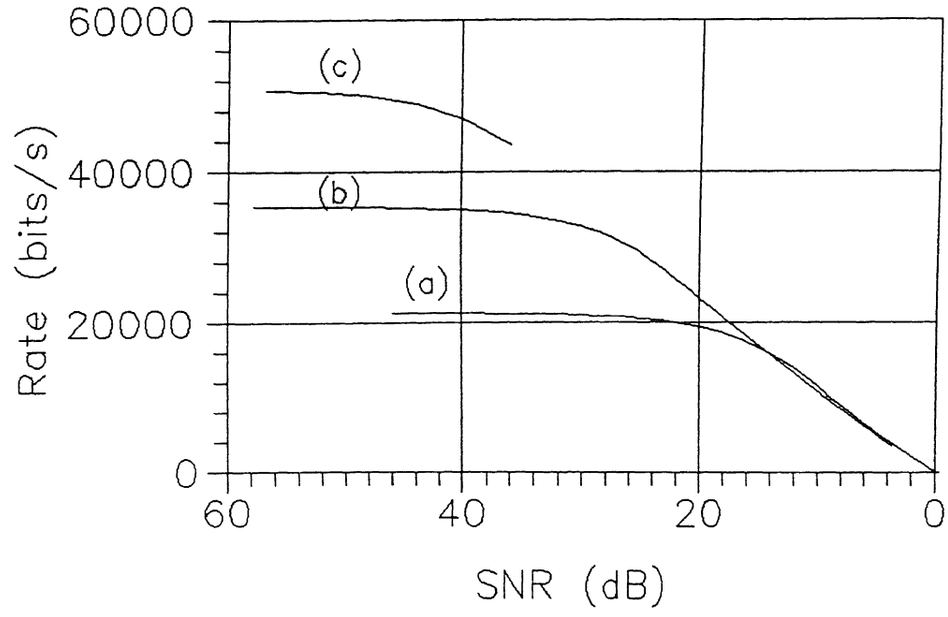


Fig. 6.5: Graph of $R_1(D)$ with respect to the mean squared error distortion criterion for quantized speech sources using Blahut's algorithm Note the logarithmic scale of the x-axis the average distortion is plotted as SNR (a) Quantization resolution = 6 bits/sample, (b) 8 bits/sample, and, (c) 10 bits/sample

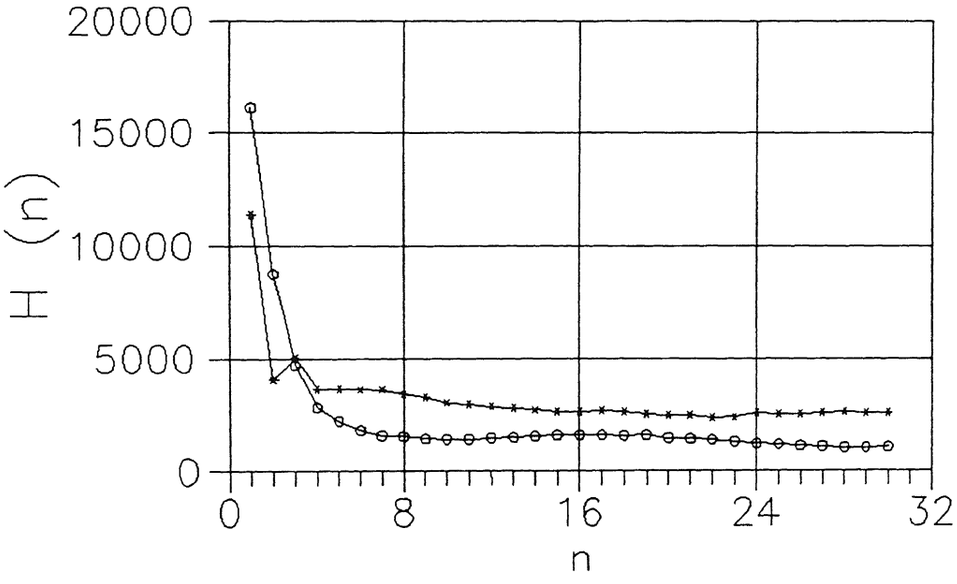


Fig 6.6: Plots of $\hat{H}_2(r,n)$ vs n showing the convergence of the entropy rate with increasing dimension (a) Female speaker Sentence utterance Why were you away a year, Roy? Data length = 20480 samples at 8 KHz, Quantization = 8 bits/sample, (b) Male speaker Sentence utterance Nanny may know my meaning Data length = 14016 samples at 8 KHz, Quantization = 10 bits/sample

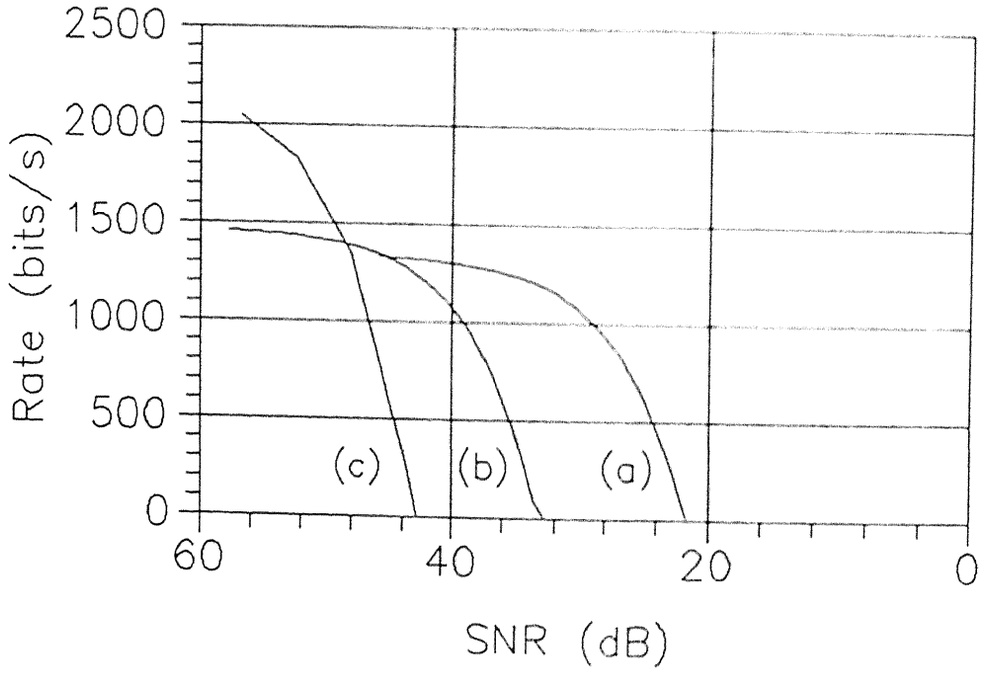


Fig. 6.7: Graph of $\tilde{R}^L(D)$, eq. (6.17a) with respect to the mean squared error criterion for quantized speech sources. The average distortion D is plotted as SNR. (a) Quantization resolution = 6 bits/sample, (b) 8 bits/sample, and, (c) 10 bits/sample.

Chapter 7

Conclusion

We have done a preliminary study of a nonlinear deterministic framework for speech signal modelling for coding applications Most medium and low bit rate speech coding schemes model the signal as an output of a time varying linear filter excited by a source. A large portion of the recent research effort has been directed to the design of appropriate excitation functions rather than investigate alternative model forms However, this research paradigm for speech coding may now be approaching a stage of saturation A major purpose of our study is to replace the ubiquitous linear filter by a more general nonlinear representational form Towards this goal, we first did a state space based analysis study of speech signals Just as a correlation analysis helps in a linear modelling exercise, the estimation of dynamical invariants such as dimension, metric entropy and Lyapunov exponents from time series is helpful in building nonlinear *deterministic* state space models of the time series We recapitulate the salient features of this analysis study and the results obtained for speech signals in the following points:

(1) The scalar time series is embedded in a *reconstructed* state space as a *reconstructed* trajectory This reconstruction is based on Takens' theorems which state that the dynamical invariants computed from the reconstructed trajectory will be the same as those of the original dynamical system whose time evolution is observed in the form of the scalar time series

(2) We have used two optimality criteria, namely the singular value decomposition method and the redundancy method to reconstruct speech trajectories, plot their 2-d projections and make observations from them

(3) Lyapunov exponents give a coordinate independent measure of the local stability properties of a state space trajectory They categorize bounded trajectories into equilibrium points, periodic solutions, quasiperiodic solutions and chaos We have computed the *largest* Lyapunov exponent of reconstructed trajectories of phoneme articulations and have compared the results with those obtained from synthetically generated periodic and quasiperiodic data From the results obtained and the comparison tests, we conclude that reconstructed trajectories can be distinguished from periodic and quasiperiodic behaviour in that nearby speech trajectories exhibit exponential divergence on the average

(4) A correlation dimension analysis of speech shows that it is largely a low dimensional signal We have also computed the correlation dimension from a simplified statistical model of a particular vowel utterance The "dimension" attribute of a time series is helpful in a state space modelling exercise because it gives the necessary and sufficient number of state space variables needed to model the data We have qualified the dimension results with a study of the various sources of error affecting the estimates

(5) Metric entropy is another dynamical invariant which quantifies the average rate of loss of information about the state of a dynamical system as it evolves in time For completely predictable systems, $K = 0$, while for chaotic systems, $0 < K < \infty$ Ideal random behaviour is characterized by $K = \infty$ The relevance of metric entropy in the context of state space modelling is that it is inversely proportional to the average time over which a dynamical system (or a time series model) can be predicted from a given initial condition We have computed the second order dynamical entropy of speech which is a lower bound of the metric entropy The positive value of the second order entropy and the largest Lyapunov exponent both give evidence of the average divergence of nearby speech trajectories in the reconstructed state space. This observation allows us to distinguish speech signals from strictly periodic or quasiperiodic behaviour

Based on the dynamical analysis results we have investigated some nonlinear prediction schemes for speech signal modelling and coding in a state space framework The results of this study are summarized in the following points:

(1) As a first choice, we have studied the (quadratic) polynomial representation form This is because it is a simple extension of a linear model and the optimal set of coefficients, in the sense of minimum m s e., can be obtained by solving a set of simultaneous linear equations The basis of comparison with short term linear prediction (LP) in terms of segmental SNR is the number of coefficients in the two predictor models We have principally considered two ordering schemes for selection of model terms of the quadratic predictor In the first method, we exhaust all possible terms upto a certain time lag before considering terms which include signal dependence upto greater lags The second method is based on orthogonal term search from a set of candidate terms While the first method does not perform as well as a LP scheme in terms of segmental prediction gain, the second method of terms arrangement gives a modest improvement over LP for the same number of coefficients It is worth noting that in a similar study of quadratic predictors [139], the basis of comparison with LP is the time delay upto which signal correlations are considered rather than the number of model coefficients In this case, the quadratic predictor performs significantly better than a LP Another reported observation in this case is that the short term quadratic predictor is also capable of modelling the pitch period redundancy to a great extent

(2) We also studied the prediction properties of a *local* prediction scheme in state space In this scheme, the representation form is optimized over a local volume in the state space where the prediction is to be done The Local State Prediction (LSP) scheme is studied in terms of segmental prediction gain, plots of the prediction error sequence, their spectrum and autocorrelation function and is compared with the error sequence resulting from (i) short term LP, and (ii) short term plus long term LP In LSP, an appropriately chosen neighbourhood will contain trajectory points that are close to the "target" vector in time as well as those which are approximately an integral number of pitch periods away Thus,

a LSP attempts to simulate the functions of both short term and long term linear prediction simultaneously The performance of a *local linear* prediction scheme can be broadly categorized as lying between a short term and short term plus long term LP in the above terms It must be noted that this comparison is unfair to the LSP because a LSP uses *previous* speech values to predict the present whereas in LP, the model coefficients are optimally estimated for a block of data and thereafter used for prediction *within* the block

(3) We also propose a framework for low to medium delay speech coding in the medium bit rate range based on LSP It is an analysis – by – synthesis coder structurally similar to CELP and named as a Vector Excited Local State Prediction (VELSP) coder The following points highlight the coding scheme and bring out the differences with CELP

(i) LSP is used instead of LP The LSP is performed using previous *reproduced* speech which is available to the decoder as well

(ii) A single excitation codebook designed from empirical data is used instead of two separate codebooks as in CELP taking advantage of the prediction property of LSP

(iii) A LSP based coder is naturally suited to low delay coding

(iv) Since a LSP based coder is *nonlinear,* we give a method to incorporate the gain factor.

We have implemented and studied a basic form of the coder structure at 5 2, 6 5 and 8.0 kb/s which require excitation sequences of lengths of 20, 16 and 13 samples respectively. While the segmental SNR performance is similar to CELP, the reproduced speech has perceptible noise and becomes poor at the 5 2 kb/s rate

We also propose an algorithm for the computation of a *lower bound* of the rate distortion function for stationary ergodic sources *with memory* Both discrete and continuous alphabet sources are considered The lower bound is more tractable than the rate distortion function itself The difficult part in computing the lower bound is in the estimation of the entropy rate / differential entropy rate using histogram technique. This is because of an exponential increase in the data length requirement

as statistical dependence for larger time frames is successively considered Specifically, we have given an algorithm for the computation of the second order entropy rate for discrete alphabet sources using the correlation sum technique which is a lower bound of its entropy rate. The algorithm is then extended to compute the second order differential entropy rate of a continuous alphabet source which is a lower bound of its differential entropy rate Examples are given to show the efficacy of the estimation scheme Finally, we compute the lower bound with respect to the mean square error distortion criterion for quantized speech sources.

Some suggestions for further work with regard to the problems studied by us are as follows

(1) In the first problem of dynamical analysis of speech signals, it will be helpful to estimate the *entire* Lyapunov exponent spectrum rather than just the largest Lyapunov exponent While the largest Lyapunov exponent is sufficient to determine the average exponential divergence of nearby trajectories in state space, a computation of the entire spectrum will provide deeper understanding of the human speech process and additionally help verify the metric entropy results

(2) One can study the suitability of other nonlinear representational forms for speech prediction and coding in terms of prediction performance and computational complexity

(3) We have implemented and studied a skeletal structure of our proposed VELSP coding scheme The two areas that require further investigation are the improvement in the quality of reproduced speech and the reduction in the computational complexity of the coder and decoder

(4) We have proposed an algorithm for the computation of a lower bound of the rate distortion function for stationary ergodic sources with memory This is based on the conjecture that the correlation sum technique will give better estimate than the histogram technique in the computation of the entropy rate This conjecture is based on the intuitive arguments presented in chapter 6 which have also been the basis for the computation of generalized dimensions and

metric entropy using the generalized correlation sum rather than the histogram technique in dynamical systems literature The veracity of these conjectures should be investigated mathematically

The theory of nonlinear dynamical analysis and deterministic state space modelling of time series has been used for various applications These include the study of fluid flows, sunspots, mechanical vibrations, economic data, climatic data and weather prediction There has been a considerable interest in the recent past to analyse and model biomedical signals such as EEG and ECG using the tools of nonlinear dynamics This intuitively appears to be an appropriate paradigm for the investigation of biomedical signals which are essentially observables of "complex" dynamical systems While several preliminary results are available, this is a potentially interesting avenue of research

Appendix A

Dynamical Systems Terminology

The purpose of this appendix is to briefly review dynamical systems terminology used in the thesis. For this, we have drawn from [34],[35],[36], [37],[38],[72] and direct the reader to these references for more elaborate discussion.

Dynamical systems are described both in terms of the real space and arbitrary manifolds In the real case, an nth-order *autonomous* dynamical system with continuous time evolution is defined by the state equation

$$\mathbf{s} = \mathbf{g}(\mathbf{s}), \quad \mathbf{s}(t_0) = \mathbf{s}_0 \tag{A 1}$$

where $\mathbf{s} \in R^n$ is the *state* at time t, and $\mathbf{g}$ $R^n \rightarrow R^n$ is called the *vector field.* $\mathbf{g}$ associates a tangent vector with each point in the state space A particular solution of (A 1) with initial condition $\mathbf{s}_0$ at time $t = t_0$ is called a *trajectory* and is denoted by $\phi(t, \mathbf{s}_0)$ The mapping ϕ $R^n \rightarrow R^n$ is called the *flow* of the system In the discrete time case, the dynamical system is defined by a map ϕ $R^n \rightarrow R^n$ and the flow is given by ϕ^n, where n indexes the discrete time

Dynamical systems can also be described in terms of manifolds Let the state space of a dynamical system be a compact manifold M. A dynamical system on M is a map ϕ $M \rightarrow M$ (discrete time) or a vector field (continuous time) The vector field associates each point in the manifold with an element of the tangent space at that point The time evolution of the dynamical system is given by the flow $\phi(t, \mathbf{s}_0)$ The

process of observing the time evolution of a dynamical system induces an *observable* which is a "smooth" function $h : M \to R$

We are interested in classifying the steady state behaviour of dynamical systems through invariants In particular, the *steady state* refers to the asymptotic behaviour as $t \to \infty$ A point p is a *limit point* of s if $\phi(t, s)$ converges to p as $t \to \infty$ The set of all limit points of s is called the *limit set*, $L(s)$, of s A set L is *invariant* under ϕ if, for all $s \in L$ and all t, $\phi(s) \in L$ Limit sets are closed and invariant under ϕ A limit set L is an *attracting limit set* or *attractor* if there exists an open neighbourhood U of L such that $\phi(t, s) \in U \; \forall \; t$ and $\phi(t, s) \to L$ as $t \to \infty$ The *domain* or *basin of attraction*, $B(L)$, is the set of all initial conditions s_0 satisfying the above definition

The classification of dynamical systems into dissipative and conservative systems comes from physics This is an extension of the notion of energy to a more general context A dynamical system in which *state space volumes* are asymptotically invariant under the dynamics is *conservative* Any dynamical system that is not conservative is *dissipative* Conservative dynamical systems do not have attractors For dissipative dynamical systems, there are two possibilities in the steady state either the trajectory is unstable and goes to infinity or it remains bounded, and approaches an attractor Nonattracting limit sets cannot be observed in physical systems or simulations

Four different types of steady state behaviour of dynamical systems are

1. **Equilibrium point or fixed point:** An *equilibrium point* s_e of a dynamical system (A 1) is a constant solution, $\phi(t, s_e)$, for all t The limit set for an equilibrium point is the point itself

2. **Periodic solution:** $\phi(t, s_p)$ is a *periodic solution* if

$$\phi(t, s_p) = \phi(t + T, s_p) \tag{A 2}$$

for all t and some minimal period $T > 0$ A periodic solution is *isolated* if $\exists$ a neighbourhood of it that contains no other periodic solution An isolated periodic solution $\phi(t, s_p)$ is called a *limit cycle*, and can occur only in nonlinear systems The limit set of a limit cycle is the closed curve traced by $\phi(t, s_p)$ over one period, and is diffeomorphic to a circle

3. Quasiperiodic solution: A *quasiperiodic solution* is a sum of periodic functions each with minimal period T_i and frequency $f_i = 1/T_i$. Also, there exists a finite set of base frequencies which are linearly independent and form a finite integer base for the frequencies f_i. A quasiperiodic solution with p base frequencies is called p-periodic. It possesses a limit set that is diffeomorphic to a p-torus.

4. Chaos: We will restrict to a qualitative discussion only. Chaos is the bounded, aperiodic solution of a dynamical system. It is a result of the twin properties of stretching and folding. *Stretching*, which is a local property, causes the exponential divergence of nearby trajectories. This is also referred to as *sensitive dependence on initial conditions*. Given two initial conditions arbitrarily close to one another, the corresponding trajectories diverge at a rate characteristic of the system. Also, since the motion is bounded in the state space, the trajectories exhibit *folding*. Thus, trajectories are locally unstable since nearby trajectories separate exponentially but at the same time, they are globally stable, since they evolve on an invariant set. The invariant limit set is not a simple geometrical object like a circle or a torus, but is related to complicated, self-similar structure called *fractals*, which are characterized by noninteger dimension. Chaotic trajectories relax to *strange attractors*. Strange attractors are attractors with the additional property of sensitive dependence on initial condition.

Appendix B

Speech Databases

The experimental results documented in the thesis are based on either of two types of speech databases The first database is based on phoneme articulations of the International Phonetic Alphabet and has been predominantly used in the dynamical analysis work of chapters 2 and 3 The second is based on phoneme specific sentence utterances and has been used for the predictive modelling, coding and rate distortion function estimation results of chapters 4, 5 and 6 We elaborate on these two databases in the following two sections

B.1 Database 1 – Phoneme Articulations

In any language, most of the sounds can be grouped into families, each family consisting of an important sound of the language together with other related sounds which represent it in particular sequences or under particular conditions of length, stress or intonation It is to such a family that the term *phoneme* is normally applied The sounds included in it are termed as its *members* or *allophones*

There are a variety of basis for the study of phonetics eg linguistics, acoustic theory of speech production etc [8], [134] In the latter category, phonemes can broadly be classified as either a continuant or noncontinuant sound Continuant sounds are produced by a fixed vocal tract configuration excited by an appropriate source, eg vowels, voiced and unvoiced fricatives, nasals etc The remaining

sounds such as plosives, stops etc. which are produced by a changing vocal tract configuration are termed as noncontinuants

The International Phonetic Alphabet (IPA), which is periodically revised, is often used for reporting experimental work based on phonemes A simple classification scheme for consonants is known as "classification by place and manner" "Place" stands in for "place of production" meaning thereby the point of articulation of the segment "Manner" is the short form for "manner of production" which primarily means the type of stricture made by the articulation process to produce the utterance The interested reader is referred to [135] for a detailed description of this classification scheme for phonemes Table B 1 lists the 57 consonants of the IPA by place and manner of articulation and 8 cardinal vowels which together form our speech database 1 Wherever two phonemes are given the same identification, the latter symbol represents a voiced articulation.

The speech database consists of the recordings of 57 consonant articulations by three male and one female speaker corresponding to the IPA chart (revised upto 1989) and cardinal vowel articulations, spoken 4 times each by one person (Daniel Jones) † Since all the phonemes cannot be spoken in isolation, one associates a *phonetic context* to mean the sounds next to it or near it in which the phoneme is a part All the consonant articulations of the database are followed by /a/ while the cardinal vowel articulations are sustained utterances Each of these utterances were lowpass filtered at 7 5 kHz and sampled at 16 kHz at 12 bit resolution The consonant articulations were isolated from the utterances by a visual inspection of the corresponding time series Since plosive utterances are transient in nature and necessarily of extremely short duration, we have excluded them from the above isolation process The 13 plosives of the IPA have therefore not been used for any analysis work reported in the thesis Thus, a total of 208 phoneme segments comprising of 44 IPA consonants spoken by 4 persons and 8 cardinal vowels spoken 4 times by one person have been used in the experiments

---

† The database was provided by the Phonetics and Spoken English Department of the Central Institute of English and Foreign Languages, Hyderabad, India.

S.No.	Phoneme Type	Symbol
1	Bilabial Plosive	p
2	Bilabial Plosive	b
3	Alveolar Plosive	t
4	Alveolar Plosive	d
5	Retroflex Plosive	ʈ
6	Retroflex Plosive	ɖ
7	Palatal Plosive	c
8	Palatal Plosive	ɟ
9	Velar Plosive	k
10	Velar Plosive	ɡ
11	Uvular Plosive	q
12	Uvular Plosive	ɢ
13	Glottal Plosive	ʔ
14	Bilabial Nasal	m
15	Labiodental Nasal	ɱ
16	Alveolar Nasal	n
17	Retroflex Nasal	ɳ
18	Palatal Nasal	ɲ
19	Velar Nasal	ŋ
20	Uvular Nasal	ɴ
21	Alveolar Trill	r
22	Uvular Trill	ʀ
23	Alveolar Tap	ɾ
24	Retroflex Flap	ɽ
25	Bilabial Fricative	ɸ
26	Bilabial Fricative	β
27	Labiodental Fricative	f
28	Labiodental Fricative	v
29	Dental Fricative	θ
30	Dental Fricative	ð
31	Alveolar Fricative	s
32	Alveolar Fricative	z
33	Postalveolar Fricative	ʃ
34	Postalveolar Fricative	ʒ
35	Retroflex Fricative	ʂ
36	Retroflex Fricative	ʐ
37	Palatal Fricative	ç

S.No.	Phoneme Type	Symbol
38	Palatal Fricative	ʝ
39	Velar Fricative	x
40	Velar Fricative	ɣ
41	Uvular Fricative	χ
42	Uvular Fricative	ʁ
43	Pharyngeal Fricative	ħ
44	Pharyngeal Fricative	ʕ
45	Glottal Fricative	h
46	Glottal Fricative	ɦ
47	Alveolar Lateral Fricative	ɬ
48	Alveolar Lateral Fricative	ɮ
49	Labiodental Approximant	ʋ
50	Alveolar Approximant	ɹ
51	Retroflex Approximant	ɻ
52	Palatal Approximant	j
53	Velar Approximant	ɰ
54	Alveolar Lateral Approximant	l
55	Retroflex Lateral Approximant	ɭ
56	Palatal Lateral Approximant	ʎ
57	Velar Lateral Approximant	ʟ
58	Cardinal Vowel	i
59	Cardinal Vowel	e
60	Cardinal Vowel	ɛ
61	Cardinal Vowel	a
62	Cardinal Vowel	ɑ
63	Cardinal Vowel	ɔ
64	Cardinal Vowel	o
65	Cardinal Vowel	u

Table B 1: List of phoneme articulations of speech database 1 The consonants are those of the International Phonetic Alphabet

B.2 Database 2 – Phoneme Specific Sentences

A set of phoneme specific sentences were proposed by Huggins and Nickerson [136] for the purpose of evaluation of the degradations of processed speech to specific phoneme types They proposed a set of four such sentences from a larger set These are as follows.

(a) Why were you away a year, Roy?

(b) Nanny may know my meaning

(c) His vicious father has seizures

(d) Which tea party did Baker go to?

These sentences are *phoneme specific* in that they concentrate all phonetic segments with similar acoustic properties in a single sentence and exclude as far as possible all phonetic segments with dissimilar properties The above sentences include among them all the consonants of English except /θ/, /l/ and /j/ The first sentence contains only vowels and glides These sounds are characterized by an all pole spectra, which change slowly, and contain no abrupt changes in the amplitude level All the sounds are voiced and only slow changes of pitch occur

The second sentence contains only (nasalized) vowels and nasals Like the first sentence, it is voiced throughout, and its spectrum and amplitude level change relatively slowly Both nasals and nasalized vowels contain zeros in their spectra

The third sentence contains only voiced and unvoiced fricatives besides vowels Fricatives contain zeros in their spectra but the spectra themselves are very different from those of voiced sounds due to the noise excitation In this sentence utterance, the rates of amplitude change are still relatively low

The fourth sentence contains (vowels and) all stops and affricates except /j/ The spectrum and amplitude of the speech waveform change frequently and abruptly, and there are many voiced–unvoiced transitions

The speech database comprises of PCM recordings of the 4 sentences spoken by 6 male and 6 female speakers † According to the information made available with

† This database was provided by the AT & T Bell Laboratories, Murray Hill, NJ, USA

the database, the recorded speech was lowpass filtered at 3.9 kHz at the time of recording. Subsequently, they were digitally filtered with an FIR filter having a flat response between 125 and 3500 Hz and a roll-off of -5 dB/100Hz above 3500 Hz. The speech signal was sampled at 8 kHz at a resolution of 16 bits/sample. The total duration of digitized speech is 120s.

Bibliography

[1] D Abercrombie, *Elements of General Phonetics,* Edinburgh University Press, 1967

[2] H Abut and N Erdol, "Bounds on $R_1(D)$ functions for speech probability models," *IEEE Trans on Information Theory,* Vol IT–25, pp 225–228, 1979

[3] B S Atal and S L Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America,* Vol 50, No 2, (Part 2), pp 637–655, 1971

[4] B S Atal and J R. Remde, "A new model of LPC excitation for producing natural sounding speech at low bit rates," *Proc Int Conf on Acoustics, Speech and Signal Process*, pp 614–617, 1982

[5] B S Atal and M R Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans on Acoustics, Speech and Signal Processing,* Vo. ASSP-27, No 3, pp 247–254, 1979

[6] R Badii, G Broggi *et al,* "Dimension increase in filtered chaotic signal," *Physical Review Letters,* Vol 60, No 11, pp 979–982, 1988

[7] C Bandt and B Pompe, "Entropy profiles of speech signals," *Physics Letters A,* Vol 175, pp 305–313, 1993

[8] M.F Barnsley, *Fractals Everywhere,* Academic, Boston, Mass , 1988

[9] W H. Bellchambers *et al,* "The International Telecommunication Union and development of worldwide telecommunications," *IEEE Commun Mag*, Vol 22, No 5, pp 72–83, 1984

[10] G Benettin, L. Galgani and J-M Strelcyn, "Kolmogorov entropy and numerical experiments," *Physical Review A*, Vol 14, pp 2338–2345, 1976

[11] N Benvenito, "The 32 kb/s ADPCM coding standard," *AT&T Tech Journal*, Vol 65, Sept./Oct 1986

[12] T Berger, *Rate Distortion Theory A Mathematical Basis for Data Compression*, Prentice Hall, Englewood Cliffs, N J 1971

[13] H-P Bernhard and G Kubin, "Speech production and chaos," *Proc 12th Int Cong of Phonetic Sciences*, France, 1991

[14] H-P Bernhard and G Kubin, "Detection of chaotic behaviour in speech signals using Fraser's mutual information algorithm," *Treizieme Colloque Gretsi*, Juan-Les-Pins Du, pp 1301–1311, Sept 1991

[15] S Bingham and M Kot, "Multidimensional trees, range searching and a correlation dimension algorithm of reduced complexity," *Physics Letters A*, Vol 140, pp 327–330, 1989

[16] R E Blahut, "Computation of channel capacity and rate – distortion functions," *IEEE Trans on Information Theory*, Vol IT-18, pp 460–473, 1972

[17] R E Blahut, *Principles and Practice of Information Theory*, Addison-Wesley Publishing Co. 1987

[18] R E Bogner and T. Li, "Pattern search prediction of speech," *Proc Int Conf on Acoustics, Speech and Signal Process*, pp 180–183, 1989

[19] W A Brock, "Distinguishing random and deterministic systems abridged version," *Journal of Economic Theory*, Vol 40, pp 168–195, 1986

[20] G Broggi, "Evaluation of dimensions and entropies of chaotic systems," *Journal of Optical Society of America B*, Vol 5, No 5, pp 1020–1028, 1988

[21] D S Broomhead and G P King, "Extracting qualitative dynamics from experimental data," *Physica D*, Vol 20, pp 217–236, 1986

[22] P Byrant, R Brown and H D I Abarbanel, "Lyapunov exponents from observed time series," *Physical Review Letters*, Vol 65, No 13, pp 1523–1526, 1990

[23] J P Campbell, V C. Welch and T E Tremain, "An expandable error – protected 4800 bps CELP coder (U S Federal standard 4800 bps voice coder)," *Proc Int Conf on Acoustics, Speech and Signal Process*, pp 735–738, 1989

[24] J G. Caputo and P. Atten, "Metric entropy An experimental means for characterizing and quantifying chaos," *Physical Review A*, Vol. 35, No 3, pp 1311–1316, 1987

[25] M Casdagli, S. Eubank *et al*, "A theory of state space reconstruction," *Information Dynamics*, eds H Atmanspacher and H Scheingraber, Plenum Press, N.Y., 1991

[26] M Casdagli, D. D Jardins *et al*, "Nonlinear modelling of chaotic time series theory and applications," *Tech Report*, Los Alamos National Laboratory, Los Alamos, N M , U S A , No LA–UR–91–1637, 1991.

[27] A Čenys and K Pyragas, "Estimation of the number of degrees of freedom from chaotic time series," *Physics Letters A*, Vol. 129, No 4, pp 227–230, 1988

[28] G J Chaitin, "Randomness and mathematical proof," *Scientific American*, Vol 232, pp 47–52, 1975

[29] J.-H Chen *et al*, "A low delay CELP coder for the CCITT 16 kb/s speech coding standard," *IEEE Journal on Sel Areas in Commun*, Vol 10, No 5, pp. 830–849, 1992

[30] H Chen, W C Kong and C C. Ko, "Comparison of pitch prediction and adaptation algorithms in forward and backward adaptive CELP systems," *IEE Proceedings – I*, Vol 140, No 4, pp 240–245, 1993.

[31] A Cohen and I Procaccia, "Computing the Kolmogorov entropy from time signals of dissipative and conservative dynamical systems", *Physical Review A*, Vol 31, No 3, pp 1872–1882, 1985

[32] T M Cover and J A Thomas, *Elements of Information Theory*, John Wiley, N Y, 1991

[33] J P Crutchfield and B S McNamara, "Equations of motion from a data series," *Complex Systems*, Vol 1, pp 417–452, 1987.

[34] R L Devaney, *An Introduction to Chaotic Dynamical Systems*, Benjamin / Cummins, 1986

[35] J-P Eckmann, S O Kamphorst *et al*, "Liapunov exponents from time series," *Physical Review A*, Vol 34, No 6, pp 4971–4979, 1986

[36] J-P Eckmann and D Ruelle, "Ergodic theory of chaos and strange attractors," *Reviews of Modern Physics*, Vol 57, No 3, Part 1, pp 617–656, 1985

[37] S Eubank and D Farmer, "An introduction to chaos and randomness," *Tech Report*, No. LA-UR 90-1874, Los Alamos National Laboratory, Los Alamos, N M, U S A, 1990, also in *1989 Lectures in Complex Systems, SFI Studies in the Sciences of Complexity, Lect Vol 2*, ed E Jen, Addison-Wesley, pp 75–185, 1990.

[38] G Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, 2nd ed, 1970

[39] J D Farmer, E Ott and J A Yorke, "The dimension of chaotic attractors," *Physica D*, Vol 7, pp 153–180, 1983

[40] J D Farmer and J J Sidorowich, "Predicting chaotic time series," *Physical Review Letters*, Vol 59, No 8, pp 845–848, 1987

[41] J D Farmer and J J Sidorowich, "Exploiting chaos to predict the future and reduce noise," in *Evolution, Learning and Cognition*, ed Y C Lee, World Scientific Press, pp 277, 1988

[42] J L Flanagan, *Speech Analysis, Synthesis and Perception*, Springer Verlag, N Y, 2nd ed., 1972

[43] J Ford, "How random is a coin toss ?," *Physics Today*, Vol 36, No 4, pp 40–47, 1983

[44] A M Fraser, "Reconstructing attractors from scalar time series A comparison of singular system and redundancy criteria," *Physica D*, Vol 34, pp 391–404, 1989

[45] A M Fraser, "Information and entropy in strange attractors," *IEEE Trans on Information Theory*, Vol 35, No 2, pp 245–262, 1989

[46] A M Fraser and H L Swinney, "Independent coordinates for strange attractors from mutual information," *Physical Review A*, Vol 33, No 2, pp 1134–1140, 1986.

[47] K Fukunaga and D R Olsen, "An algorithm for finding intrinsic dimensionality of data," *IEEE Trans on Computers*, Vol C-20, No 2, pp 176–183, 1971

[48] G Gabor and Z Gyorfi, "On the higher order distribution of speech signals," *IEEE Trans on Acoustics, Speech and Signal Process*, Vol. ASSP-36, No 4, pp 602–603, 1988

[49] R G Gallager, *Information Theory and Reliable Communication*, Wiley, N Y, 1968

[50] N Gershenfeld, "An experimentalist's introduction to the observation of dynamical systems," *Directions in Chaos, Vol 2*, ed H. B Lin, World Scientific, pp 301–379, 1988.

[51] A. Gersho, "Advances in speech and audio compression," *Proc of the IEEE,* Vol 82, No 6, pp 900–918, 1994

[52] A. Gersho and R M Gray, *Vector Quantization and Signal Compression,* Kluwer Academic Press, Boston, 1992

[53] I. Gerson and M. A Jaisuk, "Vector sum excited linear prediction (VSELP) speech coding at 8 kbps," *Proc Int Conf on Acoustics, Speech and Signal Process*, pp 461–464, 1990

[54] P. Grassberger, "Generalized dimensions of strange attractors," *Physics Letters A,* Vol. 97, pp 227–230, 1983

[55] P. Grassberger, "Finite sample corrections to entropy and dimension estimates," *Physics Letters A,* Vol 128, No 6,7, pp 369–373, 1988

[56] P. Grassberger, "An optimized box–assisted algorithm for fractal dimensions," *Physics Letters A,* Vol 148, No 1-2, pp 63–71, 1990

[57] P Grassberger and I Procaccia, "Characterization of strange attractors," *Physical Review Letters,* Vol 50, pp 346–349, 1983

[58] P Grassberger and I Procaccia, "Measuring the strangeness of strange attractors," *Physica D,* Vol. 9, pp 189–208, 1983

[59] P Grassberger and I Procaccia, "Estimation of the Kolmogorov entropy from a chaotic signal," *Physical Review A,* Vol 28, No 4, pp 2591–2593, 1983

[60] P Grassberger and I Procaccia, "Dimensions and entropies of strange attractors from a fluctuating dynamics approach," *Physica D,* Vol 13, pp 34–54, 1984.

[61] P Grassberger, T. Schreiber and C Schaffrath, "Nonlinear time sequence analysis," *Int Journal of Birfucations and Chaos,* Vol 1, No 3, pp 521–547, 1991

[62] H S Greenside, A Wolf, *et al,* "Impracticality of a box-counting algorithm for calculating the dimensionality of strange attractors", *Physical Review A,* Vol 25, No 6, pp 3453–3456, 1982

[63] J Guckenheimer and P Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields,* Springer, New York, 1983

[64] V Haggan, S M Heravi and M. B Priestley, "A study of the application of state – dependent models in nonlinear time series analysis," *Journal of Time Series Analysis,* Vol 5, No 2, pp 69–102, 1984

[65] T. Hediger, A. Passamante and M E Farrel, "Characterizing attractors using local intrinsic dimensions calculated by singular-value decomposition and information – theoretic criteria," *Physical Review A*, Vol 41, No 10, pp 5325–5332, 1990

[66] H G E Hentschel and I Procaccia, "The infinite number of generalized dimensions of fractals and strange attractors," *Physica D*, Vol 8, pp 435-444, 1983

[67] H Herzel, "Bifurcations and chaos in voice signals," *Applied Mech Rev*, Vol 46, No 7, pp 399–413, 1993.

[68] U Heute, "Medium-rate speech coding – trial of a review," *Speech Commun*, Vol 7, No 2, pp 125–149, 1988

[69] A W F Huggins and R S Nickerson, "Speech quality evaluation using 'phoneme-specific' sentences," *Journal of Acoustical Society of America*, Vol 77, No 5, pp 1896–1906, 1985

[70] W Hurwicz and H Wallman, *Dimension Theory*, Princeton University Press, Princeton, 1949

[71] A N Ince, "Speech processing standards," in *Digital Speech Processing Speech Coding, Synthesis and Recognition*, ed A N Ince, Kluwer Academic, Boston, 1992

[72] S. H Isabelle, A V Oppenheim and G W Wornell, "Effects of convolution on chaotic signals," *Proc Int Conf on Acoustics, Speech and Signal Process.*, Vo IV, pp 133–136, 1992

[73] K Ishizaka and J L Flanagan, "Synthesis of voiced sounds from a two – mass model of the vocal cords," *Bell Syst Tech Journal*, Vol 51, pp 1233–1268, 1972

[74] N Jayant, "Signal Compression Technology targets and research directions," *IEEE Trans on Sel Areas in Commun*, Vol 10, No 5, pp 796–818, 1992

[75] N S Jayant and P Noll, *Digital Coding of Waveforms Principles and Applications to Speech and Video*, Prentice Hall, Englewood Cliffs, N J, 1984

[76] D Jones, *The Phoneme Its Nature and Use*, W Heffer and Sons, Cambridge, 2nd ed, 1962

[77] H Kalverman and P Meissner, "Rate distortion bounds for speech waveforms based on Itakura–Saito segmentation," *Signal Processing IV Theories and Applications*, EURASIP, pp 127–140, 1988.

[78] J L. Kaplan and J. A Yorke, "Chaotic behaviour of multidimensional difference equations," in *Functional Differential Equations and Approximations of Fixed Points*, eds. H. O Peitgen and H O. Walther, Vol 730 of Lect. Notes in Mathematics, Springer-Verlag, Berlin, pp 204–227, 1979

[79] N. Kitawaki, M Honda and K Itoh, "Speech – quality assessment methods for speech – coding systems," *IEEE Commun Mag*, Vol. 22, No 10, pp 26–32, 1984

[80] N Kitawaki and H Nagabuchi, "Quality – assessment of speech coding and speech synthesis systems," *IEEE Commun Mag*, Vol. 26, No 10, pp 36–44, 1988

[81] W B Kleijn, D. J Krasinski and R H Ketchum, "An efficient stochastically excited linear predictive coding algorithm for high quality low bit rate transmission of speech," *Speech Commun*, Vol 7, pp 305–316, 1988

[82] M J Korenberg and L D Paarmann, "Orthogonal approaches to time-series analysis and system identification," *IEEE Signal Processing Magazine*, Vol. 8, No 3, pp 29–43, 1991

[83] Y A. Kravtsov, "Randomness, determinateness and predictability," *Sov Phys Usp*, Vol 32, No 5, pp 434–449, 1989

[84] P Kroon and B Atal, "Quantization procedures for the excitation in CELP coders," *Proc Int Conf Acoustics, Speech and Signal Process*, pp 1649–1652, 1987

[85] P Kroon and B S Atal, "On the use of pitch predictors with high temporal resolution," *IEEE Trans on Signal Processing*, Vol. 39, No 3, pp. 733–735, 1991

[86] P Kroon and B S Atal, "Predictive coding of speech using analysis – by – synthesis techniques," in *Advances in Speech Signal Processing*, eds S Furui and M M Sondhi, Marcel Dekker, N Y, 1992

[87] P Kroon and E F. Deprettere, "A class of analysis – by – synthesis predictive coders for high quality speech coding at rates between 4 8 and 16 kbits/s," *IEEE Journal on Sel Areas in Commun*, Vo. 6, No 2, pp 353–363, 1988

[88] P Kroon, E F. Deprettere and R. J Sluyter, "Regular – pulse excitation A novel approach to effective and efficient multipulse coding of speech," *IEEE Trans on Acoustics, Speech and Signal Proc*, Vol ASSP-34, pp 1054–1063, Oct 1986

[89] G. Kubin and W B Kleijn, "Time – scale modification of speech based on a nonlinear oscillator model," *Proc Int Conf on Acoustics, Speech and Signal Proc*, 1994

[90] A Kumar, *Role of Deterministic Chaos in Speech Signal Modelling*, Dept of Electrical Engineering, I I T. Kanpur, India, M Tech Thesis, No 90/216, MT, March 1990

[91] A Kumar and S K Mullick, "Application of deterministic chaos to signal modelling," *Proc of Workshop on Signal Processing, Communications and Networking*, I I Sc, Bangalore, India, pp 21–30, July 1990

[92] A. Kumar and S K Mullick, "Attractor dimension, entropy and modelling of speech time-series," *Electronics Letters*, Vol 26, No 21, pp 1790–1792, 1990

[93] A Kumar and S K Mullick, "Speech signal modelling *a la* chaos," *1990 IEEE Digital Signal Proc Workshop*, paper no 1 8, New Paltz, N Y, U S A, 1990

[94] W Liebert, K. Pawelzik and H G Schuster, "Optimal embeddings of chaotic attractors from topological considerations," *Europhysics Letters*, Vol 14, No 6, pp 521–526, 1991

[95] W Liebert and H G Schuster, "Proper choice of the time delay for the analysis of chaotic time series," *Physics Letters A*, Vol 142, No 2,3, pp 107–111, 1989

[96] J Makhoul, "Linear prediction A tutorial review," *Proc of the IEEE*, Vol 63, No 4, pp 561–580, 1975

[97] B Mandelbrot, *Fractal Geometry of Nature*, W H Freeman, N Y, 1982

[98] P Maragos, "Fractal aspects of speech signals dimension and interpolation," *Proc Int Conf on Acoustics, Speech and Signal Process*, pp 417–420, 1991

[99] J D Markel and A H Gray, *Linear Prediction of Speech*, Springer Verlag, N Y, U S A, 1976.

[100] R A McDonald, "Signal-to-noise and idle channel performance of differential pulse code modulation systems – Particular applications to voice signals," *Bell Syst Tech Journal*, Vol 45, pp. 1123–1151, 1966

[101] A I Mees, P E Rapp and L S Jennings, "Singular value decomposition and embedding dimension," *Physical Review A*, Vol 36, No 1, pp 340–346, 1987

[102] M Moller, W Lange *et al*, "Errors from digitizing and noise in estimating attractor dimensions," *Physics Letters A*, Vol. 138, No 4,5, pp 176–182, 1989

[103] F C Moon, *Chaotic and Fractal Dynamics An Introduction for Applied Scientists and Engineers*, Wiley, 1992

[104] M A H Nerenberg and C Essex, "Correlation dimension and systematic geometric effects," *Physical Review A*, Vol 42, No 12, pp 7065–7074, 1990

[105] A. R. Osborne and A Provenzale, "Finite correlation dimension for stochastic systems with power-law spectra," *Physica D*, Vol 35, pp 357, 1989.

[106] V I Oseledec, "A multiplicative ergodic theorem Liapunov characteristic numbers for dynamical systems," *Trans Moscow Math Soc*, Vol. 19, pp. 197–231, 1968.

[107] N H Packard, J-P Crutchfield *et al*, "Geometry from a time series," *Physical Review Letters*, Vol. 45, No 9, pp 712–716, 1980

[108] T S Parker and L. O Chua, "Chaos A tutorial for engineers," *Proc of the IEEE*, Vol 75, No. 8, pp 982–1008

[109] A Parker and J O Hamblen, "Optimality equations for an all-pole model with multiple impulse excitation," *Signal Processing*, Vol. 17, pp 119–127, 1989

[110] K Pawelzik and H G Schuster, "Generalized dimensions and entropies from a measured time-series," *Physical Review A*, Vol 35, No 1, pp 481–484, 1987

[111] M B Priestley, *Nonlinear and Nonstationary Time-Series Analysis*, Academic Press, London, 1988

[112] I Procaccia, "The static and dynamic invariants that characterize chaos and the relations between them in theory and experiments," *Physica Scripta*, Vol T9, pp 40–46, 1985.

[113] T F Quatieri and E M Hofstetter, "Short-time signal representation by nonlinear difference equations," *Proc Int Conf on Acoustics, Speech and Signal Process*, pp 1551–1554, 1990

[114] L R Rabiner and R W Schafer, *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs, N J 1978

[115] R P Ramachandran and P Kabal, "Stability and performance analysis of pitch filters in speech coders," *IEEE Trans on Acoustics, Speech and Signal Processing*, Vol ASSP-35, No 7, pp 937–946, 1987

[116] A Renyi, *Probability Theory*, North Holland, Amsterdam, 1971

[117] D. Ruelle, "Chaotic evolution and strange attractors," *Cambridge University Press*, 1989

[118] D Ruelle, "Deterministic chaos the science and the fiction," *Proc Royal Soc of London*, Vol A427, pp 241–248, 1990

[119] M. Sano and Y. Sawada, "Measurement of the Lyapunov spectrum from a chaotic time series," *Physical Review Letters*, Vol. 55, No. 10, pp 1082–1085, 1985

[120] J Schoentgen, "Non-linear signal representation and its application to the modelling of the glottal waveform," *Speech Commun*, Vol 9, No 3, pp 189–201, 1990

[121] M Schroeder, *Fractals, chaos, power laws Minutes from an infinite paradise*, W H Freeman, N Y, 1991

[122] M. R Schroeder and B S. Atal, "Rate distortion theory and predictive coding," *Proc ICASSP*, pp 201–204, 1981

[123] M R Schroeder and B S Atal, "Code excited linear prediction (CELP) High quality speech at very low bit rates," *Proc Int Conf on Acoustics, Speech and Signal Process*, pp. 937–940, 1985

[124] J Schroeter and M M Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing*, eds S Furui and M M Sondhi, Marcel Dekker, N Y, pp 231–268, 1992

[125] H G. Schuster, *Deterministic Chaos An Introduction*, 2nd ed, Weinheim VCH, 1989

[126] C E Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, Vol 27, pp 379–423, 623–656, 1948

[127] C E Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE National Convention Record*, Part 4, pp 142–163, 1959

[128] R. Shaw, "Strange attractors, chaotic behaviour and information flow," *Z Naturforsch*, Vol 36a, pp 80–112, 1981

[129] A. C Singer, G W Wornell and A V Oppenheim, "Codebook prediction A nonlinear signal modelling paradigm," *Proc Int Conf on Acoustics, Speech and Signal Process*, Vol 5, pp 325–328, 1992

[130] L Smith, "Intrinsic limits on dimension calculations," *Physics Letters A*, Vol 133, No. 6, pp 283–288, 1988

[131] N Sugamura and F Itakura, "Speech analysis and synthesis methods developed at ECL in NTT – from LPC to LSP," *Speech Communication*, Vol 5, pp 199–215, 1986

[132] W Szlenk, *An Introduction to the Theory of Smooth Dynamical Systems*, John Wiley and PWN–Polish Scientific Publishers, Warszawa, 1984

[133] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence*, Warwick, 1980 Lect. Notes in Mathematics, eds D A Rand and L S Young, Vo 898, Springer, Berlin, pp 366–381, 1981

[134] J. Theiler, "Spurious dimension from correlation algorithms applied to limited time-series data," *Physical Review A*, Vol 34, No. 3, pp 2427–2432, 1986

[135] J Theiler, "Efficient algorithm for estimating the correlation dimension from a set of discrete points," *Physical Review A*, Vol 36, No 9, pp. 4456–4462, 1987

[136] J Theiler, "Lacunarity in a best estimator of fractal dimension," *Physics Letters A*, Vol 133, No 4,5, pp. 195–200, 1988.

[137] J Theiler, "Estimating fractal dimension," *Journal of Optical Society of America A*, Vol 7, No 6, pp 1055–1073, 1990

[138] J Theiler, "Statistical precision of dimension estimators," *Physical Review A*, Vol 41, No 6, pp 3038–3051, 1990

[139] J Thyssen, H. Nielsen and S D Hansen, "Non-linear short – term prediction in speech coding," *Proc Int Conf Acoustics, Speech and Signal Process*, Vol 1, pp 185–188, 1994

[140] N. Tishby, "A dynamical systems approach to speech processing," *Proc Int Conf on Acoustics, Speech and Signal Process*, pp. 365–368, 1990

[141] R Togneri, M D Adler and Y Attikiouzel, "Dimension and structure of the speech space," *IEE Proceedings – I*, Vol 139, No. 2, pp 123–127, 1992

[142] H. Tong, *Nonlinear Time–Series A Dynamical System Approach*, Clarendon Press, Oxford, 1990

[143] B Townshend, "Nonlinear prediction of speech," *Proc Int Conf on Acoustics, Speech and Signal Process*, pp 425–428, 1991

[144] B Townshend, "Nonlinear prediction of speech signals," in *Nonlinear Modelling and Forecasting*, SFI Studies in the Sciences of Complexity, eds M Casdagli and S. Eubank, Addison–Wesley, pp 1–21, 1991

[145] T E Tremain, "The government standard linear predictive coding algorithm LPC–10," *Speech Technol*, pp 40–49, Apr 1982

[146] P Vary *et al*, "A regular pulse excited linear predictive code," *Speech Commun*, Vol 7, No 2, pp 209–215, 1988.

[147] J A Vastano and E J Kostelich, "Comparison of algorithms for determining Lyapunov exponents from experimental data," in *Dimensions and Entropies in Chaotic Systems Quantification of Complex Behaviour*, ed. G Mayer–Kress, Springer Verlag, N Y, pp 100–107, 1989

[148] S Wang, E. Paksoy and A Gersho, "Performance of nonlinear prediction of speech," *Proc Int Conf Spoken Language*, Kobe, Japan, pp 29–32, Nov 1990

[149] A Wolf, J. B Swift *et al*, "Determining Lyapunov exponents from a time series," *Physica D*, Vol. 16, pp 285–317, 1985

[150] L Wu and M Niranjan, "On the design of nonlinear speech predictors with recurrent nets," *Proc Int Conf on Acoustics, Speech and Signal Process*, Vol II, pp 529–532, 1994

[151] A D Wyner and J. Ziv, "Bounds on the rate distortion function for stationary sources with memory," *IEEE Trans on Information Theory*, Vol IT-17, No 5, 1971

[152] L S Young, "Entropy, Lyapunov exponents and Hausdorff dimension in differentiable dynamical systems," *IEEE Trans on Circuits and Systems*, Vol CAS-30, No 8, pp 599–607, 1983

Errata

page	position	instead of	read
ix	line 1	has been has been	has been
204	Fig.6.3, y-axis	$H(r,n)$	$\hat{H}_2(r,n)$
204	Fig.6.4, y-axis	$h(r,n)$	$\hat{h}_2(r,n)$
207	Fig.6.6, y-axis	$H(n)$	$\hat{H}_2(r,n)$
215	line 2	[34][35][36] [37][38][72]	[63][34][125] [36][108][50]
218	line 15	[8], [134]	[114], [76]
219	line 9	[135]	[1]